Application Cover Sheet (Attachment A)

Thoughtworks response

7th October, 2025

thoughtworks | COMMONWEALTH OF MASSACHUSETTS | MASSACHUSETTS TECHNOLOGY COLLABORATIVE

Attachment A

Application Cover Sheet

<table>
<tr><td colspan="4">Name of Respondent: Thoughtworks Federal, LLC</td></tr>
<tr><td>Mailing Address:
200 E. Randolph Street
25th Floor</td><td>City/Town:
Chicago</td><td>State:
IL</td><td>Zip Code:
60601</td></tr>
<tr><td>Telephone:
312-373-1000</td><td colspan="2">Fax:
312-373-1001</td><td>Web Address:
https://www.thoughtworks.com/</td></tr>
<tr><td colspan="2">Primary Contact for Clarification:
Mike OBrien</td><td colspan="2">Primary Contact E-mail Address:
mike.obrien@thoughtworks.com</td></tr>
<tr><td colspan="2">Authorized Signatory:
Viji Bommaraju</td><td colspan="2">Authorized Signatory E-mail Address:
viji.bommaraju@thoughtworks.com</td></tr>
<tr><td colspan="2">Legal Status/Jurisdiction(e.g., a Massachusetts Corporation, LLC, LLP, etc.)
Commonwealth of Virginia Corporation</td><td colspan="2">Respondents W-9: Please refer to Thoughtworks’ W-9 provided as a separate attachment.
Respondents UEI No.: PF85EA9UJGT3</td></tr>
</table>

Mike O'Brien

Client Principal

mike.obrien@thoughtworks.com
www.Thoughtworks.com

© 2025 Thoughtworks Commercial in Confidence

thoughtworks | MASSACHUSETTS TECHNOLOGY COLLABORATIVE

Authorized Application Signature and Acceptance Form (Attachment B)

Thoughtworks response

7th October, 2025

Attachment B

Massachusetts Technology Collaborative
Authorized Respondent's Signature and Acceptance Form

The undersigned is a duly authorized representative of the Respondent listed below. The Respondent has read and understands the RFP requirements. The Respondent acknowledges that all of the terms and conditions of the RFP are mandatory, and that Respondent's response is compliant with such requirements.

The Respondent understands that, if selected by the Mass Tech Collaborative, the Respondent and the Mass Tech Collaborative will execute an Agreement specifying the mutual requirements of participation.

Respondent agrees that the entire bid response will remain valid for sixty (60) days from receipt by the Mass Tech Collaborative.

I certify that Respondent is in compliance with all corporate filing requirements and State tax laws.

I certify that the statements made in this response to the RFP, including all attachments and exhibits, are true and correct to the best of my knowledge.

I acknowledge and understand the procedures for handling confidential materials submitted to the Mass Tech Collaborative as set forth in Section 3.1(d); I agree to be bound by those procedures and agree that the Mass Tech Collaborative shall not be liable under any circumstances for the disclosure of any materials submitted to the Mass Tech Collaborative pursuant to this RFP or upon the Respondent's selection.

Respondent:

Mike O'Brien

By:

(Signature of Authorized Representative)

Name:

Title:

Date:

Mike O'Brien

Client Principal

mike.obrien@thoughtworks.com
www.Thoughtworks.com

© 2025 Thoughtworks Commercial in Confidence

Massachusetts Technology Collaborative

Thoughtworks response

7th October, 2025

© 2025 Thoughtworks Commercial in Confidence

Table of contents

1. Executive summary **3**

2. About Thoughtworks: About Thoughtworks: A Strategic Partnership for Massachusetts' Responsible AI Future **5**

Our Foundational Methodology: De-Risking Investment and Delivering Public Value 6

Responsible AI by Design: From Ethical Principles to Engineered Practice 8

An Open and Evolvable Foundation for Future Innovation 9

An Unparalleled Ecosystem of Technology Leaders 10

3. Thoughtworks Proposed Solution **12**

Product-led approach 12

Technical architecture 13

4. Program plan **32**

1. Initiation & Design (team formation, architecture, governance design, dataset curation) - 6 weeks 33

2. Pilot Delivery and Ongoing Discovery (timebox tbd) 34

3. Rollout & Adoption (training, onboarding, public launch) 34

4. Post-Pilot Product Development and Scale 35

Expected Responsibilities / Team Engagement 35

5. TW team member profile **37**

6. Relevant experience & References **43**

7. Thoughtworks partnerships and accelerators **48**

Databricks as a unified data foundation 48

Cloud platform as core infrastructure and AI platform (AWS, Azure, or GCP) 49

8. Assumptions **51**

1. Executive summary

The Commonwealth of Massachusetts has set a bold vision - to be a global leader in applied, ethical, and equitable artificial intelligence. Central to that vision is the creation of the Data Commons Collaborative (DCC), a foundational platform that will expand access to high-quality data, accelerate research, and drive innovation across sectors. In order to realize this vision, The Commonwealth is looking for a technology partner who can help design and deliver a national model for a responsible AI ecosystem and Thoughtworks is well positioned to be that partner.

For over thirty years, Thoughtworks has been at the forefront of digital transformation, bringing modern software practices and product-thinking approaches to complex challenges in both the public and private sectors. We combine deep technical expertise with a commitment to human-centered design, ensuring solutions are built not only for functionality, but for adoption, sustainability, and long-term impact. Our experience with data platforms, open-source tools, and responsible AI practices positions us uniquely to deliver the Data Commons Collaborative platform (DCC) in alignment with the Commonwealth's vision.

Our approach will balance the scale and security of high-performance computing with the flexibility of cloud-based services. We will leverage the Massachusetts Green High-Performance Computing Center (MGHPCC) for sensitive workloads while building user-facing services such as the DCC website, catalog, and APIs on cloud infrastructure to maximize accessibility and scalability. Our solution will also allow seamless integration with other state AI capabilities such as AI Compute Resource (AICR) being implemented by MGHPCC and Snowflake (MA state solution) along with any necessary open-source and academic tools, including CKAN, Dataverse, and SDV. This ensures an extensible, interoperable, and future-ready foundation free from vendor lock-in.

A commitment to responsible AI will be at the core of our work. From bias detection and fairness auditing, to privacy-preserving synthetic data, to fine-grained security and governance, we will ensure that ethical principles are embedded in the platform itself. Our solution will make policy enforceable in practice through automated data lineage, tiered access controls, and comprehensive auditability. By doing so, we will help the Commonwealth establish the DCC as a trusted exemplar for the responsible use of AI in the United States.

We understand the importance of delivery discipline. Our agile, iterative methodology reduces risk, speeds value delivery, and ensures continuous feedback from researchers, startups, and industry users across the Commonwealth. This will result in a platform that launches strong and adapts over time to meet the evolving needs of its community.

Thoughtworks is ready to partner with the Massachusetts Technology Collaborative to design, deliver, and operationalize a world-class Data Commons Collaborative. We bring

a proven team, a track record of building complex, secure, and user-focused platforms, and a deep commitment to ethical technology. Together, we can help Massachusetts not only realize its vision, but set the standard for responsible AI innovation nationwide.

2. About Thoughtworks: A Strategic Partnership for Massachusetts' Responsible AI Future

The Massachusetts AI Hub Data Commons Collaborative (DCC) is a visionary initiative that requires more than a technology implementer; it requires a strategic partner capable of navigating complex technical, ethical, and community challenges to help the Commonwealth realize its goal of becoming a global leader in applied, ethical, and equitable AI. Thoughtworks is uniquely qualified to be that partner. For over 30 years, we have been at the forefront of the trends that define modern software and data engineering, helping the world's most ambitious organizations solve their most complex challenges.

Our proposal is built on the vision of a true partnership. We offer a unique combination of world-class digital product engineering, a foundational commitment to responsible technology, and a proven, product-led methodology. This integrated approach ensures the DCC will be more than a data repository; it will be a sustainable, evolvable, and trusted national model for a state-level AI ecosystem.

The Commonwealth's significant emphasis across multiple RFP sections on "ethical, responsible, and equitable" AI signals that this initiative is viewed not merely as a technical project, but as one with profound societal impact. A successful partner must therefore demonstrate more than technical compliance; they must show a foundational alignment with these principles. Our approach is built on three core differentiators that directly address this mandate and de-risk the Commonwealth's investment, as summarized in the table below.

Our Differentiator	Benefit to the Commonwealth	Direct RFP Alignment
Product-Led, Not Project-Driven	De-risks investment, ensures user adoption, and delivers tangible value faster by focusing on the validated needs of the DCC's diverse user community.	Need for a sustainable, extensible platform that serves a diverse user base including researchers, developers, and state agencies.
Responsible AI by Design	Moves beyond compliance checklists to embed fairness, transparency, and accountability into the platform's core architecture, ensuring the DCC is a trusted public asset.	High-priority mandate for AI Fairness (Section 11.0), Governance (Section 12.0), and Ethical Guidelines (Section 14.0).

Our Differentiator	Benefit to the Commonwealth	Direct RFP Alignment
Modern, Evolvable, & Open Architecture	Delivers a flexible, cloud-agnostic platform that avoids vendor lock-in, leverages open-source tools, and creates a resilient, future-proof public asset.	Requirement for a scalable, hybrid technology stack and a clear desire to build an open, flexible system that avoids proprietary lock-in.

The following sections detail how these differentiators translate into a comprehensive and low-risk approach to delivering a world-class Data Commons Collaborative for the Commonwealth of Massachusetts.

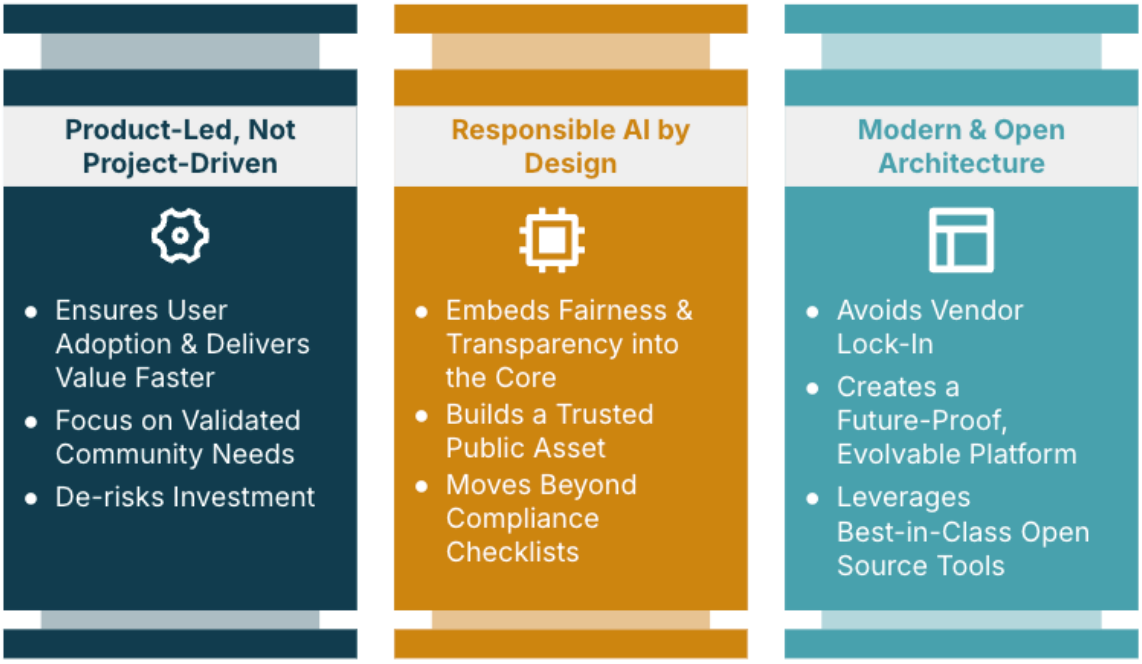

Product-Led, Not Project-Driven
Ensures User Adoption & Delivers Value Faster
Focus on Validated Community Needs
De-risks Investment
Responsible AI by Design
Embeds Fairness & Transparency into the Core
Builds a Trusted Public Asset
Moves Beyond Compliance Checklists
Modern & Open Architecture
Avoids Vendor Lock-In
Creates a Future-Proof, Evolvable Platform
Leverages Best-in-Class Open Source Tools

Figure 1: *Thoughtworks' partnership de-risks the DCC initiative by combining a product-led methodology, responsible AI by design, and a modern, open architecture to deliver a sustainable, trusted, and future-proof platform.*

Our Foundational Methodology: De-Risking Investment and Delivering Public Value

Thoughtworks' approach is designed to answer the fundamental challenge of large-scale public sector technology projects: how to deliver a complex platform that is both technically robust and genuinely useful to its intended audience. The greatest risk in an initiative of this scale is not technical failure, but building the wrong thing, a platform that is powerful but ultimately unused, leading to mission failure. Our methodology directly addresses this primary risk by embedding real users into the development process from day one, providing a powerful narrative of fiscal responsibility and mission assurance.

We differentiate our methodology from the rigid, waterfall approaches of traditional government contractors by focusing on two foundational pillars: **Product Thinking** and **Human-Centered Design (HCD).**

Product-Led, Not Project-Driven: The Product Thinking Framework

For a project of the DCC's ambition and complexity, a high-risk "big bang" approach is ill-advised. Such projects, common in government IT, attempt to define all requirements upfront, leading to brittle systems that fail to meet the evolving needs of their users. In contrast, Thoughtworks will apply its disciplined, product-led methodology to mitigate risk and ensure the platform meets the evolving needs of its diverse user community.

Our end-to-end Product Thinking framework is a system that combines Human-Centered Design, Lean-Agile Development, and a focus on measurable outcomes to build better products. This framework is institutionalized in our Product Thinking Playbook, an evolving repository of over 150 tactics and techniques drawn from Design Thinking, Agile, and Jobs-To-Be-Done Theory, used by our practitioners globally.

Instead of a "big bang" release, we will apply this framework to deliver a Minimum Viable Platform (MVP) focused on a specific, high-impact pilot community. This iterative methodology allows the joint MassTech-Thoughtworks team to:

1. **Validate assumptions and prioritize features** based on the real-world, demonstrated needs of an engaged user base, rather than abstract requirements documents.
2. **Deliver tangible value quickly** to a core group of researchers/developers, building momentum and generating early success stories that encourage broader adoption.
3. **De-risk the Commonwealth's investment** by avoiding significant expenditure on features or architectural components before they have been proven to solve a real problem for the community.

This shifts the focus from "Did we build what the document said?" to the question that truly matters: "Did we build something that people actually use and that delivers public value?"

Customer-centricity: Building for People, Not Specifications

A core tenet of our approach is that technology must be built for people and their jobs to be done. This practice ensures systems are intuitive, accessible, and solve real-world problems for their intended users. This philosophy directly counters a common failure mode in large government IT projects, where systems are built to meet technical and bureaucratic requirements but fail in user desirability and usability, leading to low adoption rates.

Our process begins with deep user research, interviews, journey mapping, and prototyping, to understand the "jobs to be done" for each of the DCC's diverse user personas, including academic researchers, AI/ML developers, and public agency staff, as identified in the RFP. This ensures the final platform is not just technically compliant but is desirable, usable, and delivers sustainable public value.

Responsible AI by Design: From Ethical Principles to Engineered Practice

The Commonwealth's repeated, high-level emphasis on "ethical, responsible, and equitable" AI is a philosophical mandate, not a technical checklist. For Thoughtworks, responsible AI is a core corporate value and a mature practice area, representing a powerful differentiator against competitors who may treat ethics as an afterthought. We move beyond mere compliance checklists to embed fairness, transparency, and accountability into the core architecture of the DCC.

Our solution operationalizes ethical AI by embedding it into the technical workflow, ensuring it is a continuous, auditable practice.

- **Bias Auditing & Mitigation:** We will integrate leading open-source toolkits like **AIF360** and **Fairlearn**, explicitly named as examples in the RFP, directly into the Databricks notebook environment. This enables a complete lifecycle approach where datasets are audited for statistical bias before training, and models are evaluated for group fairness after training. If disparities are found, developers can apply mitigation algorithms and use MLflow to create a transparent, side-by-side comparison of the baseline and mitigated models.
- **Transparency and Explainability:** The framework will incorporate tools like SHAP to help diagnose the causes of biased outcomes by explaining model behavior. All fairness assessments, mitigation steps, and SHAP visualizations are logged as MLflow artifacts, creating a fully reproducible and auditable record of every ethical intervention.

This commitment to responsible technology is not a reaction to this RFP, but a long-standing, proactive part of our corporate DNA. Our leadership in this field is demonstrated by:

- **United Nations Partnership:** Thoughtworks collaborated directly with the United Nations to co-create the Responsible Technology Playbook, which provides practical, actionable guidance for UN teams worldwide on adopting responsible practices for emerging technologies, including Generative AI.
- **MIT Technology Review Research:** We sponsored and collaborated on "The State of Responsible Technology," a global research report demonstrating our commitment to advancing the field and understanding its practical challenges.

- **Dedicated AI Governance Offering:** Thoughtworks maintains dedicated "AI Governance Strategy" and "LLM Evaluations" services, which provide clients with structured frameworks for implementing bias detection, auditing, and mitigation.

A partnership with Thoughtworks is not just the procurement of a feature; it is an alliance with a globally recognized leader in ethical technology practice, which directly de-risks the Commonwealth's core mission and reputational concerns.

An Open and Evolvable Foundation for Future Innovation

A partnership with Thoughtworks provides MassTech with more than just delivery expertise; it provides access to a culture of technological pioneering. We do not just follow best practices, we create them. The Commonwealth's technical requirements signal a sophisticated understanding of the risks of proprietary, monolithic systems, and an explicit desire for a flexible, open, and long-lasting platform. Our history and philosophy are not just compatible with this vision; we helped create it.

Pioneering Data Mesh

The revolutionary concept of **Data Mesh** was conceived and evangelized by a Thoughtworks leader. Data Mesh proposes a decentralized, domain-oriented approach to data architecture that treats data as a product This paradigm directly supports the DCC's mission to create a federated ecosystem of discoverable, high-quality, and governed data products. We have literally written the book on the topic, including foundational articles, the definitive book Data Mesh: Delivering Data-Driven Value at Scale, and practical guides based on over 75 client implementations. This positions us not just as responders to the Commonwealth's vision, but as the originators of the architectural pattern best suited to achieve it.

A Deep and Abiding Commitment to Open Source

The DCC's specified technology stack, blending platforms like CKAN and Dataverse with Python libraries like AIF360 and Fairlearn, signals a clear desire to build an open system that avoids vendor lock-in. This aligns perfectly with Thoughtworks' foundational philosophy. We are consistently rated as a top-25 corporate contributor to GitHub, having created and/or made significant contributions to dozens of influential open-source projects that are now industry standards, including Selenium (the de facto standard for web automation testing) and GoCD (a leading continuous delivery tool).

Building Evolvable and Sustainable Platforms

Government agencies rightly fear being locked into brittle, expensive systems that quickly become obsolete. Thoughtworks' core philosophy is to build evolvable, resilient platforms designed for change. This principle is embedded in our "Platform Engineering" services,

which focus on creating scalable, modular, and extensible architectures. Our proposed architecture for the DCC is cloud-agnostic, engineered to be deployed on AWS, GCP or Azure, a critical advantage given the RFP Q&A confirms there is no preferred cloud service provider. This ensures a best-fit solution for the Commonwealth and creates a truly future-proof public asset.

An Unparalleled Ecosystem of Technology Leaders

A partnership with Thoughtworks is a partnership with the entire modern technology ecosystem. We augment our world-class internal expertise with strategic, top-tier relationships with the world's leading technology providers. For a government evaluator, this multi-faceted ecosystem is a significant risk-reduction factor. It proves that our capabilities are validated by the technology leaders themselves and that we can bring the right expertise and resources to bear, regardless of the final technology choices.

Strategic Multi-Cloud Partnerships

Thoughtworks is a premier-level partner with all three major cloud service providers (CSPs), a unique capability that ensures our architectural recommendations are unbiased and centered on the best-fit solution for the Commonwealth's specific needs.

- **Amazon Web Services (AWS):** We are an AWS Premier Partner holding over 15 advanced competencies, including the crucial **Government Competency** and the new **Generative AI Competency**.
- **Google Cloud (GCP):** We are a GCP Premier Partner with specializations in Application Development and Data & Analytics. Our partnership is further distinguished by our recognition as the **2023 Google Cloud Partner of the Year for Diversity, Equity, and Inclusion**, an award that strongly resonates with the DCC's public sector values and ethical mandate. We are also among a select group of Google Partners to be awarded the **Google Public Sector Delivery Competency Badge for AI/ML**.
- **Microsoft Azure:** As a key Microsoft partner, we have deep certification paths and delivery experience across the Azure platform, including its advanced services for Data and AI.

Strategic Data & AI Partnership: Databricks

Our partnership with Databricks, the data and AI company, is a cornerstone of our data modernization strategy. We are targeting Elite tier status, the highest level, with a global team of over 175 Databricks-certified professionals. Our joint value proposition combines Thoughtworks' leadership in Data Mesh and agile engineering with the power of the Databricks Lakehouse Platform to deliver measurable impact. Our joint success stories include building a complete MLOps platform for a U.S. biopharmaceutical company using

Databricks and Delta Lake, demonstrating our ability to execute on the very architecture we propose for the DCC.

By combining our proven, product-led methodology, a foundational commitment to responsible AI, deep expertise in open and evolvable platforms, and a world-class partner ecosystem, Thoughtworks will ensure the DCC is not only delivered successfully but becomes a sustainable, trusted, and powerful engine for innovation for the Commonwealth of Massachusetts.

3. Thoughtworks Proposed Solution

Product-led approach

To ensure the success of the DCC, Thoughworks has set out a northstar architectural approach that supports the full aspirations represented by the strategic requirements set out in the RFP.

However, in order to launch valuable capabilities and tools to the innovation community as quickly as possible, Thoughtworks proposes a product-thinking-led approach to build a targeted pilot focusing on the needs of a specific segment of the intended community as a way to drive design and prioritization decisions. By selecting a targeted community with similar jobs-to-be-done and data needs, the team will be able to transparently:

- Prioritize types of data based on the needs of the pilot community
- Estimate usage/storage/capacity needs
- Run active design and solution feedback sessions with an engaged user base
- Prioritize admin and support features based on numbers of expected pilot users - limits the risk of over-investing before proving value

Taking the pilot approach allows Thoughtworks to provide a realistic timeline and outcome-oriented roadmap to get valuable features to market to an engaged and excited pilot community.

Taking a customer-oriented and outcome-oriented approach is core to Thoughtworks ways of working. Our end-to-end Product Thinking framework combines the best practices of Human Centered Design and Lean-Agile Development to build better products that benefit users and drive business value.

Product thinking follows four key principles:

- **Customer Centricity:** grounding decisions in the validated needs of the DCC community (initially the pilot)
- **Outcome Orientation:** success measured by active experiments being run by the pilot community
- **Experimentation & Learning:** validating functionality through prototypes, pilots, and iteration.
- **Cross-Functional Collaboration:** embedding product, design, engineering, and SME roles into durable, multi-disciplinary teams.

To ensure we continue to design and build the most important data tools and features that meet user needs, drive objectives, and work within the constraints of the surrounding technical ecosystem, we proactively address four product risks, desirability, usability, feasibility, and viability, throughout development.

Thoughtworks will ensure a pilot-user centered discovery period drives priorities for an initial pilot release. From that point forward, the team engages in continuous discovery as they deliver ensuring each set of features is validated for desirability (does this solution address an important problem or job-to-be-done) and usability (does the solution work for the users in their context).

Technical architecture

We are committed to the Massachusetts AI Hub's preference for open-source technologies. Our solution leverages industry-leading managed services like **Databricks** and native **Cloud Platform services** (AWS/Azure/GCP) to provide the necessary security, scale, and compliance overhead, while our core functional capabilities such as data discovery, synthetic data generation, and ethical AI are delivered through **MLflow** and other best-of-breed open-source frameworks (**AIF360, Fairlearn, CKAN**). This solution is designed to not only meet but exceed the requirements set forth in the RFP, establishing a world-class data commons that will serve as a catalyst for AI innovation across the Commonwealth.

Our vision is to deliver a singular, cohesive ecosystem that centralizes data access, fosters collaboration, and champions ethical AI development. By leveraging the Databricks Lakehouse architecture, the solution seamlessly combines the performance, governance, and reliability of a traditional data warehouse with the flexibility and scale of a data lake. This unified approach eliminates data silos, streamlines workflows for researchers and developers, and provides a robust foundation for the DCC's long-term growth and success.

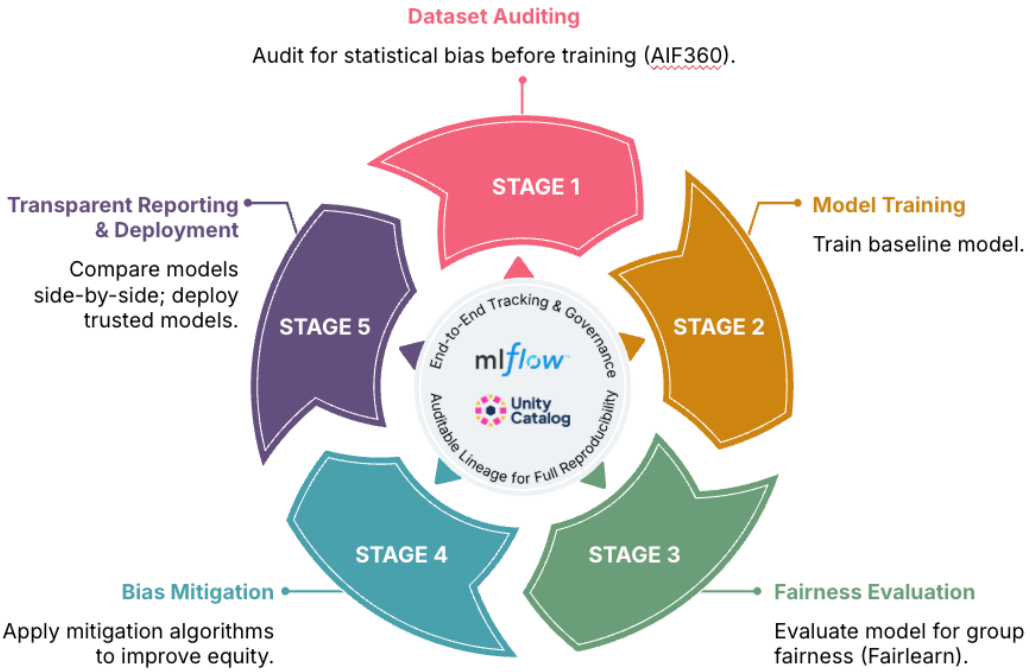


Figure 2: *We operationalize ethical AI by embedding fairness checks, bias mitigation, and full reproducibility directly into the ML lifecycle, tracked and governed by MLflow and Unity Catalog.*

The platform will empower users - from academic researchers and AI/ML developers to startups and state government agencies - to discover, access, and utilize a rich catalog of curated datasets. Key capabilities, including privacy-preserving synthetic data generation, robust data curation and versioning, and an integrated AI fairness and bias toolkit, are not treated as separate add-ons but as deeply integrated components of the platform. Governance and security are woven into the fabric of the architecture through Databricks Unity Catalog, ensuring that policies are defined once and enforced everywhere, from data ingestion to model deployment.

The following diagram provides a high-level overview of the proposed architecture, illustrating the end-to-end flow of data and the interaction between the core components of the DCC platform.

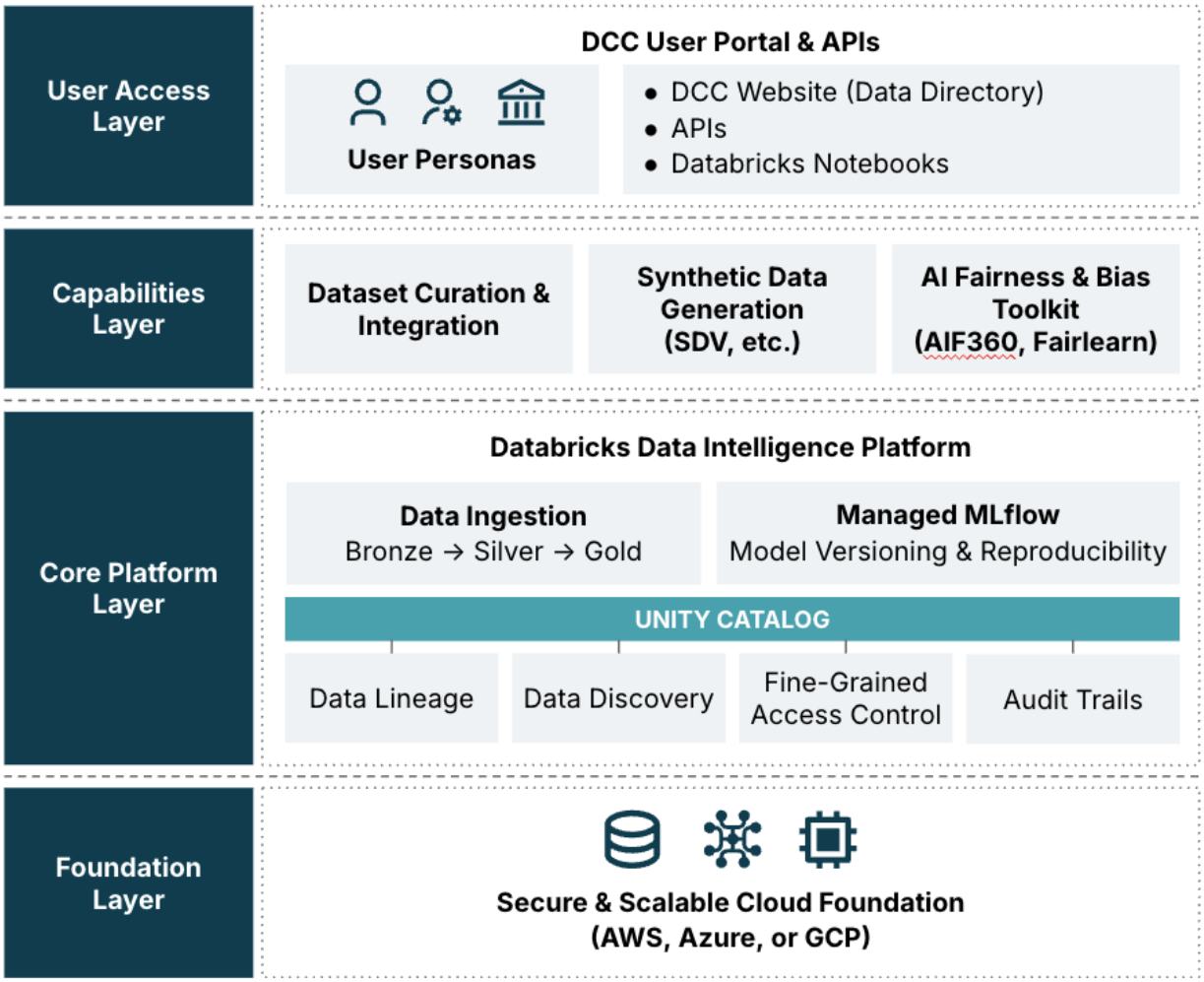
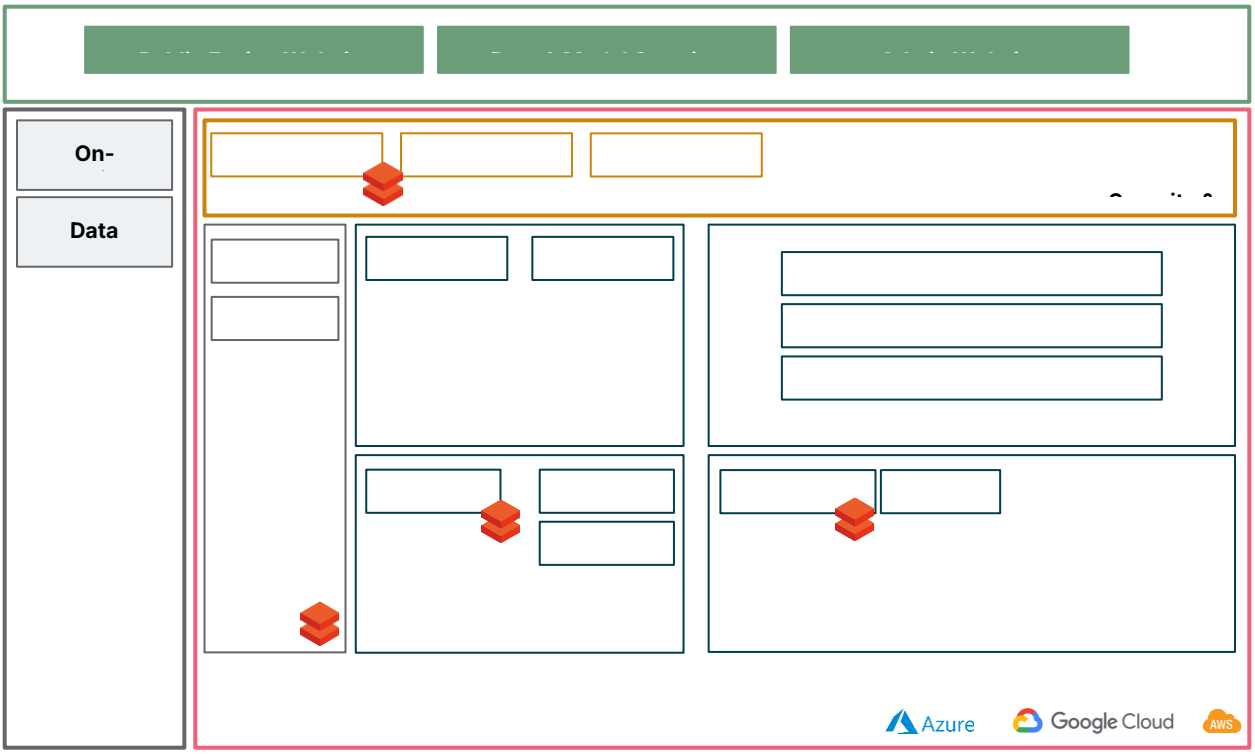


Figure 3: *Our proposed architecture leverages the Databricks Data Intelligence Platform on a secure cloud foundation, creating a unified, scalable, and governable ecosystem for AI innovation.*



This architecture is inherently modular and extensible, designed to evolve with the needs of the Massachusetts AI ecosystem. It provides the extensible foundation required for future expansion, including user-generated dataset contributions and collaborative research initiatives, ensuring the DCC remains at the forefront of AI innovation for years to come.

Data commons website

The DCC website will serve as the primary digital gateway for the entire Massachusetts AI ecosystem, providing a centralized, intuitive, and powerful portal for data discovery, access, and collaboration. The design philosophy is centered on creating a seamless and integrated user experience that caters to the diverse needs of its target users, from AI/ML developers and academic researchers to state agencies and startups. The website will be fully responsive, ensuring a consistent and accessible experience across desktop, tablet, and mobile devices.

User Experience and Interface Design

The user journey is designed to be intuitive and efficient. Upon arrival, users will be greeted with a clean, modern homepage that provides an introduction to the DCC, highlights featured dataset collections (Public, Massachusetts, Synthetic), and offers a prominent quick search bar. A section for recent updates will keep the community informed of new datasets and platform capabilities.

The core of the discovery experience lies in the **Dataset Directory**. This will be a searchable and filterable list of all available data assets. Users can employ a powerful faceted search to filter datasets by critical metadata attributes, including:

- **Data Type:** Text, Image, Video, Tabular, GIS
- **Source:** Public/Global, Massachusetts State, Synthetic Tools
- **Domain:** Healthcare, Environment, Transportation
- **File Format:** CSV, GeoJSON, JSON, Parquet
- **License:** Open, Academic Use, etc.

Each dataset in the directory will be represented by a **Dataset Card**, providing at-a-glance information such as its name, a brief description, relevant tags, and its source (e.g., Hugging Face, MassGIS). From the card, users can navigate to a **Dataset Detail Page,** which provides comprehensive metadata, including data size, full description, license details, categories, links to external documentation, and a list of related datasets.

Core Functionality: Discovery, Exploration, and Contribution

Beyond discovery, the website will facilitate deeper data exploration and community engagement. For hosted datasets, a **Dataset Explorer** will allow users to preview data directly in the browser without needing to download it. This functionality will include viewing sample rows of tabular data and inspecting the dataset's schema, fulfilling a key user experience enhancement requirement.

To foster a collaborative environment, the website will feature a **Community Dataset Submission** form. This will allow users to suggest new external data links or propose datasets for hosting, capturing valuable feedback and expanding the commons organically.

Administrative functions will be managed through a secure, role-based **Admin Dashboard**. This dashboard will empower authorized curators and administrators to add, edit, or remove datasets and links, manage user submissions, and monitor the health of external links through an integrated validation dashboard.

The website will also serve as the user-facing portal for other key DCC functions. Users will be able to request access to restricted datasets directly through the portal and manage their "Commons Credits" by viewing their balance and transaction history in an integrated section of their user profile. This unified approach avoids a fragmented user experience, ensuring that all interactions with the DCC such as data/model discovery, access and credit management all occur within a single, cohesive interface.

Meeting Performance, Accessibility, and Security Mandates

Our evolutionary architecture approach ensures that the platform is designed to not only meet the necessary non-functional requirements as per the RFP but also be able to embrace change in future and evolve with business needs.

- **Performance:** To achieve the sub-2-second page load and search performance targets, the frontend will be built using a modern JavaScript framework (such as React or Vue.js). Static assets will be served via a Content Delivery Network (AWS CloudFront or Google Cloud CDN) to ensure low latency for users globally. Backend APIs will be optimized for speed and efficiency.
- **Accessibility:** The website will be developed to meet the accessibility standards set by MassTech, adhering to Web Content Accessibility Guidelines (WCAG) 2.1 Level AA. This includes using semantic HTML, providing proper ARIA (Accessible Rich Internet Applications) attributes, ensuring keyboard navigability, and conducting both automated and manual accessibility testing throughout the development lifecycle.
- **Security:** All traffic will be encrypted using HTTPS. To prevent common web vulnerabilities, all user inputs will be rigorously sanitized on both the client and server sides. To protect against denial-of-service attacks and ensure fair usage, API endpoints will be protected with rate limiting.

Common Credit System

To incentivize contributions, promote efficient resource utilization, and sustain engagement, the DCC will feature a sophisticated Commons Credits System. This system will be built on a secure, serverless, and cryptographically verifiable architecture, directly addressing all requirements outlined in Section 7.0 of the RFP. The architecture is designed for scalability, auditability, and seamless integration with the broader DCC platform, while remaining cloud-agnostic.

Architecture: A Serverless and Verifiable Ledger

The core of the credit system will be a centralized ledger implemented using a managed database service that provides cryptographic verifiability. The RFP's requirement for a "secure, auditable ledger" and its mention of "future blockchain integration" highlight the need for data integrity guarantees like immutability and cryptographic proof. While a full blockchain introduces unnecessary complexity and performance overhead for a centrally managed system, modern cloud databases now offer built-in ledger-like capabilities that provide these benefits without the drawbacks.

Our approach leverages more sustainable, mainstream, and cloud data services that deliver the required immutability and verifiability:

- For an AWS deployment, the recommended approach is to use Amazon Aurora PostgreSQL. Configured with the pgAudit extension, PostgreSQL features database triggers that write all changes to a separate, append-only history table. This architecture, recommended by AWS as the migration path from QLDB (discontinued now) , provides a robust and fully auditable transaction log.

- For a Google Cloud (GCP) deployment, Google Cloud Spanner is an excellent choice. As a globally distributed, strongly consistent relational database, it is purpose-built for applications like financial ledgers that require absolute transactional integrity and scalability. Its inherent consistency guarantees serve as a strong foundation for an auditable system of record.

The business logic for all credit operations will be encapsulated in **serverless functions** (e.g., AWS Lambda, Google Cloud Functions), exposed via a secure REST API managed by a **managed API gateway** (e.g., Amazon API Gateway, Google Cloud API Gateway). This serverless architecture ensures that the system scales automatically to meet demand and operates on a pay-per-use model, providing exceptional cost-effectiveness and modularity. The API will provide endpoints for all required operations, such as earning, spending, and querying credit balances.

Transactional Logic: Earning and Spending Credits

The system will implement clear and automated rules for the issuance and redemption of Commons Credits, as required by the RFP.

- **Earning Credits:** Serverless functions will handle the automated issuance of credits based on defined events.
  - **Contributions:** When a user's dataset submission is approved and ingested, an event will trigger a function to issue a predefined number of credits to their account.
  - **Grants:** An administrative endpoint will allow authorized MA AI Hub personnel to grant credits to specific users or organizations.
  - **Purchase:** The system will integrate with a payment gateway to allow users to purchase credits with real currency.
- **Spending Credits:** Credits will be the currency for accessing premium resources within the DCC.
  - **Premium Access:** Accessing a premium dataset or model will trigger a transaction to debit the user's account.
  - **Compute Usage:** Deploying or training models on the DCC's compute infrastructure will consume credits based on resource utilization.
  - **Third-Party Services:** Unlocking applications or services from third-party providers on the platform will also be managed via credit transactions.

Every transaction, whether an issuance, transfer, or redemption, will be recorded in the ledger. Each record will include comprehensive metadata, a timestamp, and a purpose tag (e.g., "data contribution," "model training"), ensuring a complete and auditable historical log for every user account. Enforcement mechanisms such as credit expiration rules and tiered access will be built into the business logic to prevent misuse.

Governance, Transparency, and User Experience

The system is designed for full transparency and ease of use. A dedicated set of administrative tools will be provided for the Governance Body to review credit issuance policies, set usage limits or quotas for users, and generate detailed usage reports and dashboards for stakeholders.

For end-users, the credit system will be seamlessly integrated into the main DCC website. Through their user profile, individuals can:

- View their current credit balance and a complete transaction history.
- Understand the rules and opportunities for earning and spending credits.
- Discover resources and applications that are eligible for credit use.

The entire system will comply with state data security and privacy standards, with all components designed to be secure and interoperable with the broader DCC architecture

Data Catalog & Metadata Management System

The foundation of a successful data commons is its ability to make data discoverable, understandable, and trustworthy. To achieve this, the solution will implement a comprehensive Data Catalog and Metadata Management System using **Databricks Unity Catalog**. Unity Catalog is not merely a feature but the central pillar of governance, security, and discovery for the entire DCC platform, providing a single source of truth for all data and AI assets. This unified approach directly addresses the disparate requirements of RFP Section 8.0 in a cohesive and powerful manner.

Unity Catalog: A Single Source of Truth for All Data Assets

Unity Catalog provides a centralized governance solution that organizes all data assets into a clear, three-level namespace: Catalog + Schema + Asset. This structure will be used to logically organize all DCC resources, including:

- **Hosted Datasets:** Stored as managed Delta tables.
- **External Datasets:** Registered as external tables pointing to data in its original location.
- **Synthetic Datasets:** Clearly tagged and stored as managed tables.
- **AI/ML Models:** Registered and versioned within the catalog.
- **Functions:** Reusable user-defined functions for data processing.

This unified view allows administrators to manage and govern all assets through a single interface, regardless of their type or physical location. Unity Catalog seamlessly manages both **managed tables**, where Databricks controls the full lifecycle of the data in the open Delta Lake format, and **external tables/volumes**, where it governs access to data

residing in cloud storage without taking ownership of it. This flexibility is crucial for integrating the wide variety of data sources required by the DCC.

Enabling FAIR Principles through Rich Metadata and Discovery

The implementation of Unity Catalog is fundamentally aligned with the FAIR data principles (Findable, Accessible, Interoperable, Reusable), a key requirement of the RFP.

- **Findable:** Unity Catalog provides a powerful search interface within the Databricks UI (Catalog Explorer) and through a REST API. Users can discover datasets, models, and other assets by searching across names, comments, and custom tags. This rich search capability is the cornerstone of data discovery.
- **Accessible:** Access control is a native function of Unity Catalog. A fine-grained, SQL-based permissions model (GRANT/REVOKE) governs who can view, query, or modify each asset. This ensures that data is accessible but only to authorized users under defined policies, as detailed further in the Security and Governance sections.
- **Interoperable:** By standardizing on the open-source Delta Lake format for managed tables, the solution ensures that data is not locked into a proprietary format. Furthermore, Unity Catalog's support for custom tagging allows for the application of standard metadata schemas like DCAT or Schema.org to enhance interoperability.
- **Reusable:** To be confidently reused, data requires rich context. Unity Catalog provides this context through comprehensive metadata, including user-defined descriptions, tags, and, most importantly, automatically captured data lineage.

Automated Lineage and Auditing for Full Transparency

Two of the most powerful features of Unity Catalog are its automated lineage and auditing capabilities, which provide unparalleled transparency and trust.

- **Data Lineage:** Unity Catalog automatically captures and visualizes data lineage at the column level for any workload, whether it is written in SQL, Python, or R. This means a user can select any table, column, or dashboard and see a complete, end-to-end graph of how it was created, from the raw source data through every transformation step. This capability is essential for debugging, impact analysis, and fulfilling the provenance requirement of the RFP.
- **Auditing:** All access to data governed by Unity Catalog is automatically logged. These detailed audit logs capture who accessed what data, when, and from where. The logs are made available through system tables, which can be easily queried to monitor for compliance, investigate incidents, and generate security reports.

RFP Requirement	Capability & Benefit	Unity Catalog Feature?

Dataset Discovery & Search	Faceted search, full-text search, and APIs enables users to quickly discover relevant datasets.	✅
Metadata Management	A complete, auditable context for every dataset supports FAIR principles and enables trust in datasets published on DCC's platform. Comments, Tags, System Tables, Automated Lineage.	✅
Catalogue Publishing & Versioning	Data deposit workflow, review process, version control allow for traceable governance of catalogue items. Applies ubiquitously to datasets, ML models, and synthetic data alike.	✅
Access Control & Governance	Role-based permissions, access tiers, and audit trails.	✅
Interoperability & Integration	REST APIs, SDKs, FAIR compliance, DOI support.	✅
Data Usage & Metrics	Download and access metrics, citation tracking. Supports investment & commons credits system.	✅
Collaboration & Community	Annotations, comments, team-based curation.	✅
Flexible Data & Models Catalogue	Data, ML Models, and any other catalogue entity can be treated as first-class, discoverable/searchable entries. This allows DCC to enable synthetic data, ML models, and any future features all within the same system, regardless of type.	✅

Dataset curation and integration

A world-class data commons is defined by the quality and relevance of its data. This section outlines a strategic and systematic plan for the curation, versioning, and lifecycle management of datasets within the DCC, addressing the requirements of RFP Section 9.0. The approach combines expert-led vetting with a robust technical framework for data ingestion and versioning, leveraging MLflow to ensure reproducibility and auditability.

Curation Criteria and Ingestion Workflow

The inclusion of any dataset into the DCC's hosted collection will be governed by a formal curation process based on clear, pre-approved criteria. As stipulated in RFP section 9.1, these criteria will ensure that all curated datasets are:

- **High-Value for AI/ML:** Directly relevant to key sectors for the Commonwealth, such as transportation, healthcare, climate, and robotics.
- **Extensively Used:** Sourced from well-established benchmark datasets widely used in the AI research community.
- **Properly Licensed:** Freely shareable under open or restricted public licenses that are compatible with the DCC's governance policies.
- **Non-Sensitive or Anonymized:** Devoid of personal or sensitive information, or deemed appropriate for the creation of synthetic derivatives.

The ingestion workflow for a new dataset will proceed through a multi-stage pipeline:

1. **Submission and Identification:** Datasets will be identified either through proactive curation by domain experts or via the community submission portal on the DCC website.
2. **Vetting and Approval:** Each proposed dataset will be reviewed by a panel of domain experts and the Data Governance Body to ensure it meets all inclusion criteria.
3. **Processing and Standardization:** Upon approval, the dataset will be processed through a reliable data pipeline built with Databricks Lakeflow. This process will include data cleansing, schema validation, and conversion to the highly-performant and reliable Delta Lake format.
4. **Registration and Documentation:** The final, curated dataset will be registered as a new table in Unity Catalog. This step includes the application of comprehensive metadata, including a detailed description, source information, licensing terms, and relevant tags.

Dataset versioning and lineage with MLflow

To meet the critical requirements for version control and data lineage , the solution will leverage

MLflow Dataset Tracking. This approach moves beyond simple file versioning and creates a direct, auditable link between data and its use in AI development.

For each curated dataset, and for every subsequent update, a new version will be logged as an MLflow Dataset object within a dedicated MLflow experiment. This process captures essential metadata that guarantees reproducibility:

- **Source:** A reference to the dataset's origin, such as the S3 path of the raw data file. This provides an unbreakable lineage chain, allowing any version of the curated data to be traced back to its source.
- **Digest:** A unique SHA-256 hash of the dataset's content. This cryptographic fingerprint serves as a definitive version identifier, ensuring that researchers can always verify they are using the exact same data as in a previous experiment.

- **Schema and Profile:** The dataset's structure and summary statistics are captured, providing quick insights into its contents.

This methodology creates an immutable, auditable history of every dataset's evolution. The true power of this approach, however, emerges from MLflow's integration with the model training lifecycle. When a developer uses a specific version of an MLflow-tracked dataset to train a model, MLflow automatically records the dataset's name and digest as part of the model training run's parameters. This creates a powerful, end-to-end lineage graph that connects a specific model version in the registry directly back to the exact dataset version used to train it. This provides a level of reproducibility far superior to systems where data versioning and model tracking are handled by separate, disconnected tools. A researcher can examine any model in the DCC and definitively identify, retrieve, and reuse the precise data that created it.

Initial dataset selection

The initial phase of the DCC will focus on curating a foundational set of high-impact datasets. The process will begin with a thorough vetting of the suggested initial datasets listed in RFP Section 9.2. This includes a diverse collection of widely-used benchmark datasets across several key domains:

- **General-Purpose and Benchmark:** ImageNet, COCO, MNIST, CIFAR-10/100.
- **Natural Language Processing (NLP):** Wikipedia Dumps, Common Crawl, The Pile.
- **Health and Life Sciences:** MIMIC-IV, PhysioNet, TCGA.
- **Autonomous Systems and Robotics:** KITTI, nuScenes.
- **Social, Demographic, and Urban Data:** U.S. Census, NYC Taxi Data.
- **Massachusetts Specific Public Datasets:** Data from MassGIS, MBTA, and the Boston Open Data Portal.

This initial collection will provide immediate value to the target user community and serve as a robust foundation for the continued growth of the Data Commons.

Synthetic Data Generation Capability

At Thoughtworks, we partnered with an **Ivy-league Research University** and a **global market research firm** to build their Synthetic Data-as-a-Service platform. This platform enables the firm’s market researchers to select from a curated selection of Synthetic Data generation ML models (Model-as-a-Service) and upload a sample dataset. Users are then able to track data generation progress, cost, and other attributes. This data, once generated, is used to power millions of surveys each year, drastically improving their speed-to-market and reliance on error-prone human surveys.

Thoughtworks acted as AI advisory SMEs and owned the end-to-end technical development of the cloud-based platform, model deployment, and API management.

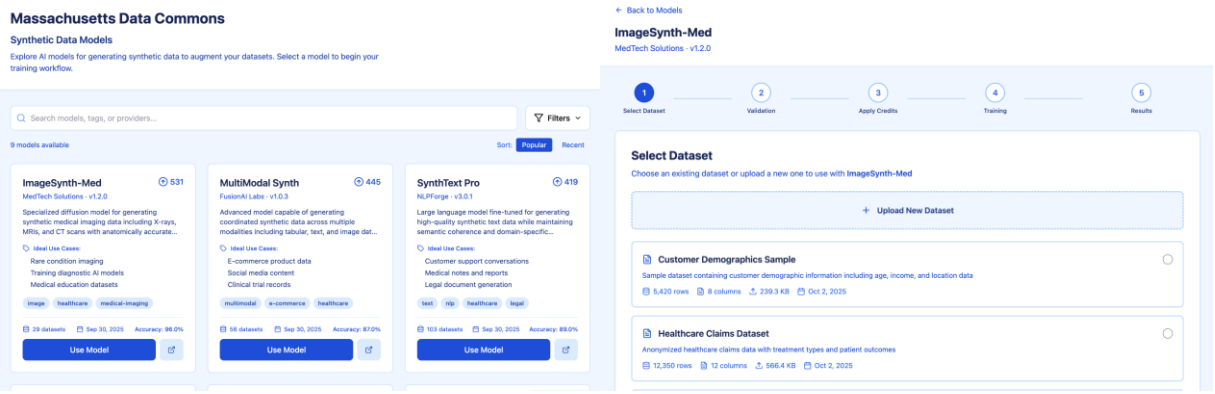
For Massachusetts DCC, synthetic data will be enabled through MLFlow's seamless integration with dataset publishing into Unity Catalogue.

A pre-approved set of curated state-of-the-art synthetic data generation models are published and made available by DCC Admin curators through Unity Catalogue and MLFlow. Models are published as cards with key information like:

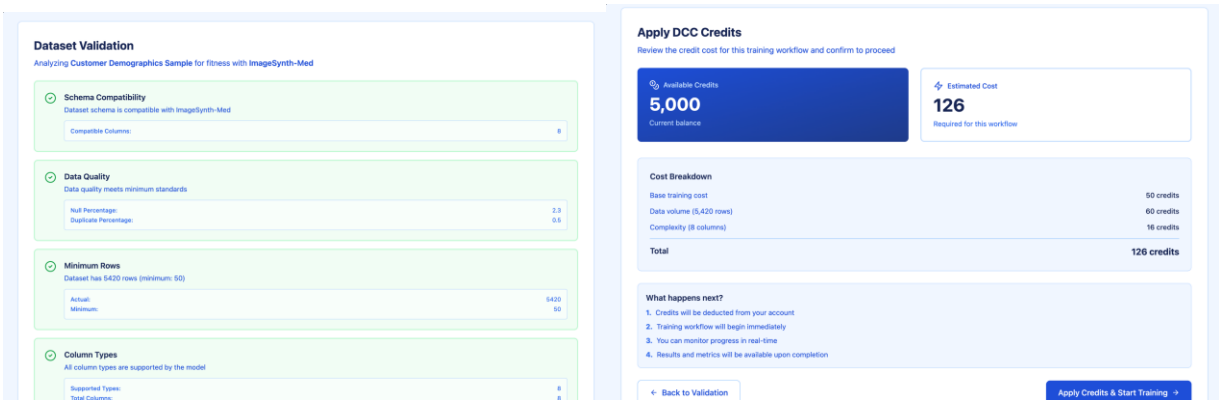
- A description of ideal use cases
- Link to algorithm methodology
- Link to search page with datasets generated by this model
- Community votes, comments, etc.

Suggestive workflow: a high-level overview of synthetic data generation

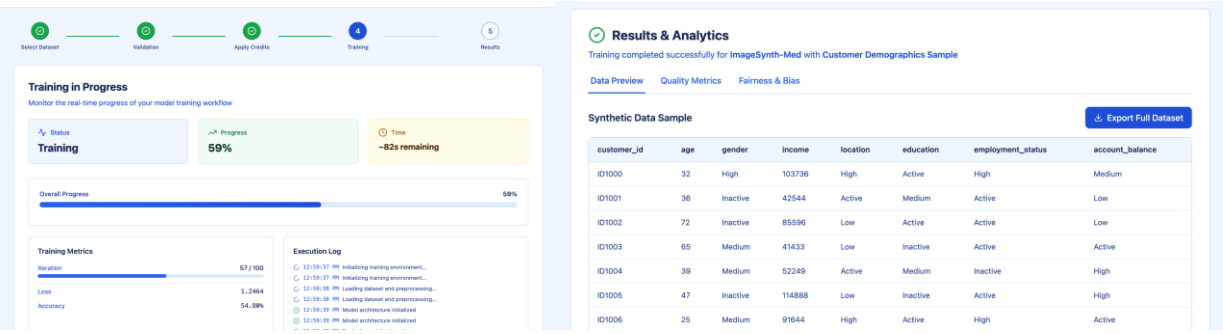
Users are presented with a list of models and when a user selects a model, users are navigated to the data generation workflow. Users will then be able to select one of their datasets or create a new dataset.



Next, datasets are analyzed for fitness against the selected model and other stagegates as determined by MTC. A user can then apply DCC credits (if required) and execute a model generation workflow.



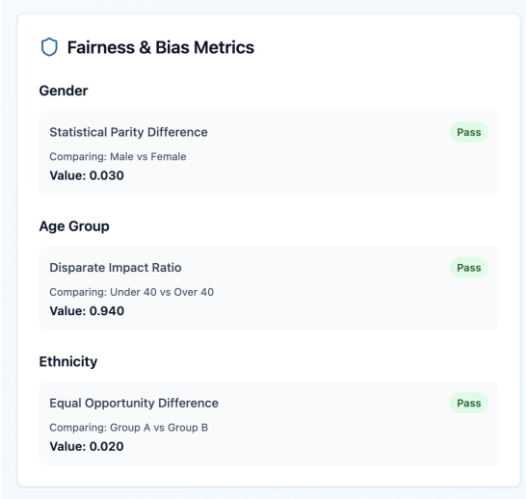
During this time, users are able to view job metrics and progress. Once training is completed, the workflow execution page is updated and users can preview the dataset to sanity check completeness & quality.



Finally, a synthetic dataset is published to the data commons catalogue. However, not without key indicators and metrics to promote AI transparency and trust in the data:

- Synthetic data marker
- Training metrics (sample dataset size, distribution, etc.) (opens to full-screen model page)
- Model metrics
- Dataset metrics (opens to full-screen model page)
- Carbon impact
- Usage metrics
- Community upvotes
- Community comments

Optionally, synthetic datasets can be queued for DCC admin review before publishing.



AI fairness and bias framework

In alignment with the DCC's guiding principles of equity and accountability, the platform will provide an integrated AI Fairness and Bias Capability. This framework is designed to empower developers and researchers to proactively identify, measure, and mitigate potential biases in datasets and AI models throughout the entire machine learning lifecycle. The solution directly addresses the requirements of RFP Sections 11.0 and 12.0 by operationalizing ethical AI principles through a practical, auditable workflow.

Integrated Fairness Tooling in the Databricks Environment

The capability will be built upon the integration of leading open-source Python libraries, as recommended by the RFP. These toolkits will be made available within the collaborative Databricks notebook environment, easily installable on compute clusters .

AI Fairness 360 (AIF360): Developed by IBM, this comprehensive toolkit provides an extensive library of over 70 fairness metrics for datasets and models, along with more than 10 bias mitigation algorithms. Its strengths lie in its breadth of metrics and its support for pre-processing (data transformation), in-processing (fairness-aware algorithms), and post-processing (outcome adjustment) mitigation techniques.

- **Fairlearn:** A Microsoft-backed project, Fairlearn offers a user-friendly API focused on group fairness assessment and mitigation. Its core component, the MetricFrame, allows for the disaggregation of any standard machine learning metric (e.g., accuracy, precision, error rate) across different subgroups defined by sensitive attributes (e.g., race, gender). This makes it exceptionally powerful for identifying allocation harms and quality-of-service harms.

The Ethical AI Lifecycle: From Data to Deployment

The fairness and bias capability is not a single tool but an integrated workflow that injects ethical checkpoints at critical stages of the ML lifecycle. Every step of this process will be tracked using **MLflow**, ensuring that all fairness assessments and mitigation efforts are reproducible, comparable, and auditable.

1. **Dataset Auditing (Pre-Processing):** Before model training begins, every curated dataset will undergo a bias audit. Using metrics from AIF360, such as the Disparate Impact Ratio and Statistical Parity Difference, the system will analyze the dataset for potential biases related to protected attributes. The resulting fairness report will be logged as an artifact to the MLflow run associated with that dataset's version, providing a baseline fairness assessment.
2. **Model Fairness Evaluation (Post-Training):** After a baseline model is trained, it will be evaluated for fairness using Fairlearn's MetricFrame. This will involve calculating key performance metrics for each subgroup. For example, in a loan application model, the system would compare the false positive rate and false negative rate across different demographic groups. These disaggregated metrics provide a clear picture of whether the model is performing equitably.
3. **Bias Mitigation and Comparison:** If significant disparities are detected, developers can apply a bias mitigation algorithm. For example, they could use a pre-processing technique like Reweighing from AIF360 to adjust the dataset or an in-processing algorithm like ExponentiatedGradient from Fairlearn to train a new model subject to fairness constraints.

4. **Tracking and Decision Making:** A new MLflow run will be initiated for the mitigated model. The same performance and fairness metrics will be calculated and logged. The MLflow UI will then provide a side-by-side comparison of the baseline model and the mitigated model, allowing developers and the Governance Body to make an informed decision by evaluating the trade-offs between overall accuracy and fairness improvements.

This entire workflow beginning all the way from audit to mitigation and comparison is captured by MLflow, creating a transparent and auditable record of every ethical intervention.

Transparency, auditability, and explainability

To further enhance transparency, the framework will incorporate model explainability tools like **SHAP (SHapley Additive exPlanations)**. SHAP can be used to explain both global model behavior and individual predictions, helping to diagnose the underlying causes of biased outcomes. SHAP summary plots and other visualizations will be logged as MLflow artifacts, providing deeper insights into model behavior.

Furthermore, **Databricks Lakehouse Monitoring** can be configured to continuously track fairness and bias metrics for models deployed in production. This system can automatically generate dashboards and trigger alerts if a model's performance begins to drift and exhibit unfair behavior over time, ensuring ongoing ethical oversight.

Security, governance and compliance

We believe in baking in security and governance controls into every aspect of the platform development lifecycle. Given our solution is primarily based on Databricks Lakehouse Platform, our approach will be to establish a robust, auditable framework using the foundational components of the Databricks Lakehouse Platform as well as the native services of the Underlying Cloud Platform (AWS, Azure or GCP). This dual-layered strategy is designed to achieve the high standards of legal compliance, ethical practice, and data integrity as mandated by the RFP.

Identity, Access Control (RBAC), and Data Protection

The entire DCC platform will be deployed within the client's own cloud account, providing full control and ownership over the underlying infrastructure. This deployment leverages core cloud security services to establish a secure foundation. We will implement a defense-in-depth model, with Unity Catalog serving as the central enforcement point for all data and AI assets.

This dual-layered security model offers a significant advantage. The Commonwealth’s central IT and security teams retain full control over the foundational cloud infrastructure (VPC, storage policies, KMS keys), satisfying enterprise security and compliance mandates. Simultaneously, the DCC's data governance team is empowered to manage all

data-level access policies dynamically and efficiently within Unity Catalog. This creates a "define once, secure everywhere" paradigm, where a policy set in Unity Catalog is consistently enforced whether the data is being accessed from a notebook, a BI tool, or an API call, thereby eliminating potential security loopholes. This comprehensive approach ensures that the DCC platform will be not only technically advanced but also fully compliant, transparent, and ethically managed from day one.

Business Capability	Technology	Approach
User Authentication (MFA/SSO)	Cloud IAM & Databricks	We will implement Multi-Factor Authentication (MFA) for all user roles and integrate with MassTech's existing Single Sign-On (SSO) service via the CSP's Identity and Access Management (IAM) solution.
Network Isolation	AWS VPC, Google Cloud VPC, etc.	The Databricks compute plane, where all data processing occurs, will be deployed within a dedicated Virtual Private Cloud (VPC). This isolates the compute resources from the public internet and allows for the configuration of strict network access control lists (ACLs) and security groups to manage traffic flow.
Role-Based Access Control (RBAC)	Unity Catalog	We will define and enforce precise, granular permissions for all required roles: Public Viewer, Researcher, Curator, Admin, Data Developer, and Auditor. Access is controlled at the table, row, and column level.
Dataset Tiering	Unity Catalog	We will implement access tiers such as Public, Restricted (requiring review/approval), and Internal, ensuring restricted datasets are protected via an automated access approval workflow integrated into the Admin Dashboard.
Data Protection	Cloud Platform & Databricks	Ensure all data is encrypted at rest (using native cloud storage encryption, such as AES-256) and in transit (via TLS 1.2+ across all APIs and user interfaces).

Auditing, monitoring, and transparency

Our proposed approach to building the platform ensures full visibility and accountability across all system activity.

- **Logging and Audit Trails:** We will utilize Databricks' Unity Catalog as well as the native Cloud Platform's logging services to capture and retain all audit logs for a minimum desirable retention period (usually three years, as per compliance best practices). This includes logging every instance of data access, dataset/model changes, Commons Credits transactions, and security events.
- **Intrusion Detection & Monitoring:** We will configure the Cloud Platform's native monitoring services to provide real-time intrusion detection and security event monitoring, ensuring platform integrity and operational resilience.
- **Transparency and Provenance:** MLflow and Unity Catalog will automatically track the full data and model lineage, enabling administrators and auditors to trace any artifact (model or dataset) back to its original source, transformations, and governance approval steps.

Governance, ethical framework, and legal compliance

We will embed ethical oversight and regulatory compliance directly into the technical workflows, rather than treating them as separate, manual steps.

- **Ethical review workflow:** We will implement a formal workflow to support ethical review of datasets and models. This includes requiring Bias Impact Statements as part of the model submission process and providing the Admin Dashboard with tools to document ethical oversight and approval.
- **Bias mitigation and auditing:** The AI Fairness and Bias Framework (built on MLflow integration) will provide statistical fairness metrics and visualizations. All fairness assessments will be automatically logged and made auditable, fulfilling the requirement for Bias Impact Documentation.
- **Terms of use and consent management:** We will define clear Terms of Use and ethical guidelines, and implement technical mechanisms (tokenization, time-bound access) for managing consent, ensuring alignment with state and federal laws (e.g., non-discrimination, anti-kickback statutes).
- **Security reviews:** We commit to defining processes for regular security reviews and penetration testing to be conducted by independent third parties, proactively maintaining the platform's security posture.

The governance policies and ethical guidelines of the DCC will be guided by a set of core principles designed to foster trust, ensure accountability, and promote equitable and responsible innovation. These principles, as required by RFP Section 14.0, will be embedded into the platform's technical architecture and operational procedures, ensuring they are not just ideals but enforceable standards. The guiding principles are as follows:

- **Transparency:** The DCC will operate with maximum transparency. All governance policies, data curation criteria, and access request procedures will be publicly documented. The platform's use of automated data lineage tracking within Unity

Catalog ensures that the provenance of every dataset and AI model is clear and auditable.

- **Accountability:** Clear lines of responsibility for the curation, security, and ethical use of data will be established. The platform's comprehensive audit logs will track all data access and system actions, ensuring that all activities can be traced back to a specific user or process, thereby enforcing accountability.
- **Fairness and equity:** The platform is fundamentally designed to mitigate bias and promote equitable access. The integrated AI Fairness and Bias Capability will provide tools to audit datasets and models for fairness issues. Furthermore, the Commons Credits System is designed to ensure equitable access to compute resources across the diverse user community.
- **Security and privacy:** A robust, multi-layered security model and a privacy-by-design approach are central to the platform's architecture. Data will be protected through encryption, network isolation, and fine-grained access controls, with a particular focus on safeguarding any sensitive or restricted datasets.
- **Sustainability:** The long-term viability and growth of the DCC will be supported by policies that encourage contributions from the community. The Commons Credits System is a key mechanism for incentivizing the contribution of high-quality datasets, models, and tools, creating a virtuous cycle of growth and innovation.

These principles will be operationalized through the technical capabilities of the Databricks platform and the formal processes of the DCC Governance Body, creating a comprehensive framework that ensures the Data Commons Collaborative serves as a responsible and powerful engine for AI advancement in Massachusetts.

4. Program plan

While the requirements document outlines five logical program phases (Initiation & Design, Infrastructure & Setup, Development, Testing & Governance, Rollout & Adoption), our recommended execution methodology prioritizes getting to market quickly and driving innovation with DCC users. Thoughtworks proposes re-mapping these phases to prioritize shipping working, usable software for feedback from pilot users.

This product-thinking approach mitigates the risk of the DCC expending budget on tools that may not fully support use cases and community users' jobs to be done. Our iterative and agile approach (specifically, Scrum and/or Kanban) helps to manage complexity, incorporate continuous stakeholder feedback, and ensure rapid, validated delivery of value.

Mapping RFP Phases to Proposed Phases in the response

RFP Phase	Response Phase
1. Initiation & Design	1. Initiation & Design (no change)
2. Infrastructure & Setup	2. Pilot Delivery and Ongoing Discovery
3. Development	2. Pilot Delivery and Ongoing Discovery
4. Testing & Governance	2. Pilot Delivery and Ongoing Discovery
5. Rollout & Adoption	3. Pilot Launch Rollout & Adoption (training, onboarding, public launch of pilot product)
n/a	4. Post-Pilot Product Development and Scale

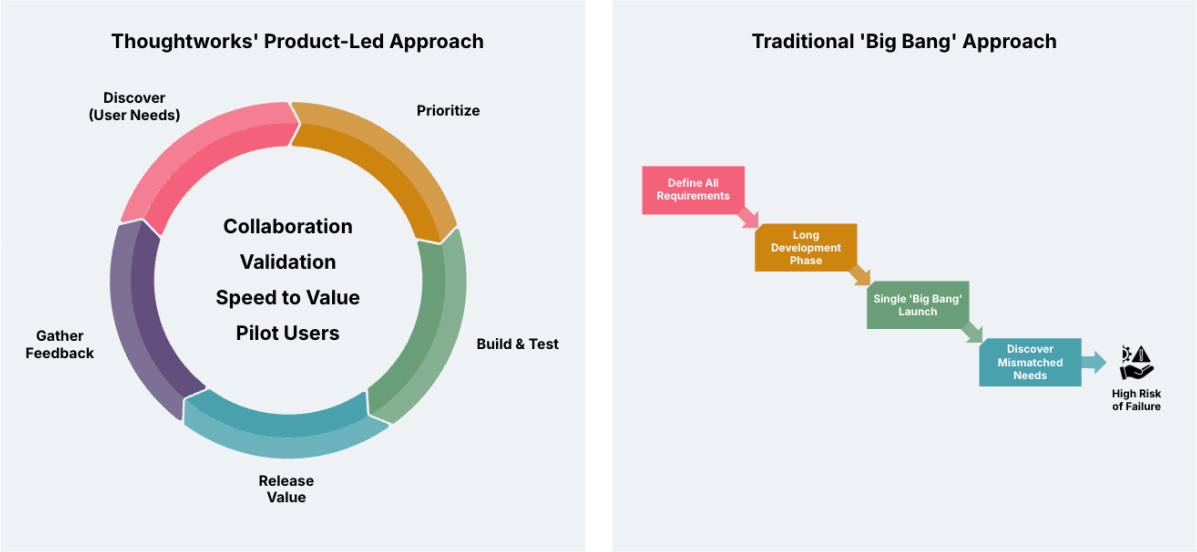


Figure 4: *Our iterative, product-led approach gets value to the community faster by focusing on continuous user feedback and discovery, mitigating the risks of a traditional "big bang" project.*

1. Initiation & Design (team formation, architecture, governance design, dataset curation) - 6 weeks

Objective: Thoughtworks team will kickoff the program of work and collaborate with the DCC to identity the target pilot audience that will allow us to:

1. Transparently prioritize datasets and feature needs
2. Validate assumptions for usage costs
3. Start experiments/prototypes with active feedback loops and get to market quickly

Key Activities and Deliverables

- Team kickoff
- Internal/external stakeholder mapping
- Program planUser-driven dataset prioritization
- Website and tool prototypes and ProofsofConcept
- Audience research, outreach and pilot audience selection
- Pilot outreach plan
- Initial Pilot release roadmap - outcomes and intended features
- Initial Pilot backlog constructed Initial pilot backlog estimated for initial release
- Validate northstar architectural and tool decisions

- Validation pilot license and run costs

2. Pilot Delivery and Ongoing Discovery

Objective: The cross-functional Thoughtworks team will work in transparent sprints against the published backlog to implement and launch working software aligned to the priorities of the target pilot audience.

Key Activities and Deliverables

- Backlog / user story grooming and ongoing prioritization
- Implementation and testing of selected tools and infrastructure
- Ongoing release of working software to production environments - validated via test driven development and automated testing
- Transparent demonstrations of working software in each sprint for MTC stakeholders and target pilot users
- Ongoing user experience testing invitations to the pilot community
- Communications / training sessions with MTC stakeholders and the pilot community
- Ongoing discovery and validation sessions with MTc stakeholders and the pilot community
- Validation of release candidates and preparation for public pilot release

3. Rollout & Adoption (training, onboarding, public launch)

Objective: Successfully launch the public version of the pilot release. For the target audience segment, the release will be fully available with valuable features and data sets, validated through pilot user feedback through delivery. Publish the future-looking roadmap to serve subsequent audience segments and additional scale.

Key Activities and Deliverables

- Prioritized features and datasets available in public-facing production environment
- Training sessions/content and launch outreach to pilot community
- Post-pilot roadmap including prioritized subsequent audience segments
- Validated post-pilot and scaling budget

4. Post-Pilot Product Development and Scale

Objective: Expand the audience of community members driving innovation through the hub

Key Activities and Deliverables

- Continuous discovery
- Identification of additional audience needs and resulting roadmap updates
- Continuous backlog construction, grooming
- Continuous build/test/ship product code
- Ongoing change management / training / support

Expected Responsibilities / Team Engagement

In order to provide context for our expected engagement, Thoughtworks offers the following expected engagement model between the Thoughtworks team and the MTC team. The following is presented as typical expectations and the team is excited to create a joint vision with MTC stakeholders.

In this case, rather than describing a RACI, we propose the following engagement levels:

- Collaborate - assumes transparent collaboration and co-design resulting in team alignment
- Recommend - assumes the Thoughtworks team will provide recommendation that the client will approve
- Approve - assumes the client will approve recommendations provided by the Thoughtworks team
- Drive - assumes the Thoughtworks team will own and drive in full transparency
- Review - assumes the client can participate or have access at any time

	MTC Program leads	MTC Technology leads	Thoughtworks team
Pilot audience definition	Collaborate	Collaborate	Collaborate
Pilot data/feature prioritization	Collaborate	Collaborate	Collaborate
Pilot outreach plan	Approve	Approve	Recommend
Pilot Spike/POC plan	Approve	Approve	Recommend
Pilot outcome roadmap	Collaborate	Collaborate	Collaborate
Pilot backlog	Review	Review	Drive
North Star Architecture	Collaborate	Collaborate	Collaborate
Pilot Technical design	Approve	Approve	Recommend

Sprint planning / backlog grooming	Review	Review	Drive
Sprint Demos	Collaborate	Collaborate	Collaborate
Pilot user outreach	Review	Review	Drive
Pilot public release candidate	Collaborate	Collaborate	Collaborate
Pilot public release planning	Collaborate	Collaborate	Collaborate

5. TW team member profile

Chris Kramer

AI Practice Lead

Chris Kramer is an AI & Machine Learning expert specializing in synthetic data and AI data architectures. He led a groundbreaking partnership between a Fortune 500 market research firm and an Ivy League research institution to build state-of-the-art digital twins and synthetic data platforms, work that fundamentally transformed how the organization approached market research by replacing costly, unreliable human respondents with synthetic alternatives.

Currently, Chris leads the Applied AI/ML Division of Thoughtworks' North America Data and AI Service Line, where he oversees AI engineering teams tackling complex data challenges across industries. With over 10 years of experience, he specializes in natural language processing, AI agents, LLMs, and knowledge graph structures that underpin modern data catalogues and AI applications.

His work has generated measurable impact: an NLP platform automating extraction across millions of PDFs (saving tens of thousands of operational hours at a Fortune 500 financial services firm) and ML applications generating millions in revenue through novel credit modeling techniques.

Chris has contributed to O'Reilly's *Fundamentals of Software Engineering and speaks regularly on generative AI and innovation.*

Arun Srinivasan

Principal Data Architect

Arun Srinivasan serves as the Principal Data Architect and AI Platforms Practice Lead for Thoughtworks, bringing deep technical leadership to the design and implementation of the Commonwealth’s Data Commons Collaborative (DCC). Arun will lead the end-to-end data architecture, ensuring the DCC’s hybrid platform is built to the highest standards of scalability, security, and governance as outlined in the RFP. His mandate is to translate

the Commonwealth's vision of a responsible, extensible, and ethical AI ecosystem into a production-grade, cloud-agnostic data platform that enables innovation without compromising trust.

Arun's expertise spans the full modern data lifecycle, from ingestion and curation through synthetic data generation, governance, and AI integration. He has architected enterprise-scale platforms on AWS and Databricks, aligning directly with the RFP's hybrid model that leverages MGHPCC/AICR for compute-intensive workloads and cloud services for accessibility and interoperability. His mastery of Databricks Lakehouse architecture, Delta Sharing, and Unity Catalog governance ensures fine-grained Role-Based Access Control (RBAC), automated data lineage, and auditable transparency, capabilities explicitly required under the DCC's governance and compliance framework.

As Thoughtworks' AI Platforms lead, Arun will ensure the DCC's architecture seamlessly integrates privacy-preserving synthetic data generation tools such as SDV and Synthea, and fairness frameworks like AIF360 and Fairlearn for model auditing. His leadership bridges data engineering and applied AI, enabling the DCC to operationalize responsible AI through embedded bias detection, explainability, and automated governance pipelines. These capabilities align directly with the RFP's ethical AI and fairness mandates.

Arun's prior work demonstrates his ability to deliver complex, high-value data systems across sectors of critical importance to the Commonwealth. For a major healthcare provider, he led Generative AI strategy and implementation, performing value analyses for clinical use cases and building production-ready data pipelines on Databricks to power AI-driven insights. For a global environmental nonprofit, he architected a data platform that supported large-scale analytics and computer vision workloads, enhancing performance, reducing onboarding time by 50%, and improving query speeds by 40%. Earlier in his career at Bloomberg Industry Group, Arun engineered a cloud-native enterprise data lake that integrated governed, multi-domain data sources using AWS Lake Formation and Glue, establishing the governance foundation for analytics and ML teams enterprise-wide.

Arun combines deep technical fluency with a pragmatic understanding of delivery. He has built and led data and AI engineering teams across both the public and private sectors, ensuring that the DCC will not only meet its initial launch milestones but evolve sustainably over time. His architectural decisions are grounded in transparency, reusability, and performance, ensuring that the Commonwealth's Data Commons Collaborative stands as a national model for responsible, ethical, and AI-ready data infrastructure.

Sharren Bates

Product Management Principal Consultant

Sharren Bates is Thoughtworks Product Management Principal Consultant and is a passionate, proven, value-driven technology strategy executive with over 25 years of experience. She serves her clients as a trusted partner offering leadership in modern, customer-oriented Product Management practice. She is a Digital Transformation subject matter expert driving time-to-value and continuous improvement.

THOUGHTWORKS EXPERIENCE 2016-present

Principal Consultant - Fortune 5 Healthcare/Retail Leader

Designed and implemented a cross-functional coaching/pairing program. Crafted learning outcomes, engagement model and curriculum modules for cross-functional team best-practices including customer/actor journeys, service design blueprinting, story and acceptance criteria writing, backlog construction and management. Designed and implemented a series of executive workshops including Enterprise Architecture Solutioning, Engineering Effectiveness Practice Leadership, Digital Design Leadership. Each leading to strengthened cross-functional executive leadership influence.

Principal Consultant - Financial Services firm with over $39 billion in retirement assets Architected a value-driven portfolio management governance structure. Built customer- and valueoriented, evidence-based governance program with the C-Suite and portfolio leaders. Improved time to-value by targeting and reducing waste and churn in decision making, resource allocation and prioritization. Led engagement turnaround with multiple cross-functional development teams. Mentored and coached Product and Technical Leads through agile discovery and inception processes - leading to scope and impact clarification. Turnaround resulted in significant increase in client satisfaction and continued partnership for additional 24 months.

Principal Consultant - Premier academic medical center in NYC

Grew cross-functional, custom software development team from 4 to a globally distributed team of 45. Built product roadmap and backlog. Drove to launch and value across three business units at the hospital system. Launched new web and mobile patient digital experiences architecting new service-oriented integration patterns with the legacy EMR and hospital HR systems. Designed, built and deployed scheduling applications for ambulatory care, telehealth and imaging services.

Marcus Gruel

Data Strategist

Marcus Gruel serves as the Lead Data Strategist for Thoughtworks, guiding the Commonwealth in defining and operationalizing the governance, data-management, and adoption frameworks that will sustain the Massachusetts AI Hub Data Commons Collaborative (DCC). He bridges business strategy and data architecture, ensuring that the DCC's structure, metadata standards, and user-access models are purpose-built to advance transparency, equity, and data-driven innovation across agencies. Marcus's leadership focuses on establishing the policies, metrics, and decision frameworks that enable responsible AI ecosystems to thrive, directly supporting Sections 12.0 (Governance Requirements) and 14.0 (Ethical Guidelines) of the RFP.

Marcus brings nearly a decade of experience shaping enterprise data-strategy and governance programs for Fortune 500 and public-sector clients. At Thoughtworks, he has led modernization and governance initiatives for clients such as ExxonMobil, Moody's Analytics, Boeing, and Komatsu Mining, applying frameworks that align regulatory compliance with business value. He has served as a strategist on multi-cloud data-platform migrations, including Microsoft Purview and AWS Lake Formation, defining operating models for metadata management, lineage, and data-quality automation. Marcus's certification as a Certified Data Management Professional (CDMP) and his hands-on proficiency with SQL, Python, and Power BI allow him to translate governance theory into actionable implementation plans, ensuring traceability and auditability from dataset ingestion through AI model use.

Before joining Thoughtworks, Marcus was a consulting leader at First San Francisco Partners, where he advised clients on data governance program maturity, collaboration models, and self-service data enablement. He developed data-quality scorecards and business-data glossaries that advanced data trust and regulatory readiness for highly regulated enterprises in insurance and financial services. Marcus is known for a listen-first approach that helps stakeholders align on value outcomes before technical implementation, an approach that resonates with the DCC's collaborative mission to foster a statewide AI innovation ecosystem. His blend of strategic governance leadership and hands-on data modernization expertise ensures the Commonwealth will stand up a Data Commons that is trusted, sustainable, and built to adapt as AI capabilities evolve.

Chloe Blanchard

Experience Design Principal Consultant

Chloe is an Experience Designer boasting an impressive background in user experience design and strategy. Her fervour for comprehending user behaviour and the intricate systems they engage with is evident in her exceptional work. Upon joining Thoughtworks, Chloe's expertise in leading exploratory, generative, and evaluative research engagements has been invaluable. Her keen eye for detail, combined with her prowess in design and prototyping, has significantly contributed to the success of numerous delivery engagements.

Bose: Led exploratory and generative research to identify product opportunities in the shared mobility space, resulting in 2 de-risked product concepts and 2 patent applications.

RBC: Led the generative and evaluative research to reimagine the social payments experience on RBC's mobile banking app. Went on to manage two designers through the design and development of the reimagined feature set.

lululemon: Led a series of lean research sprints to evaluate innovative e-commerce concepts, providing actionable insights and recommendations to their Futures Innovation & UX/Product teams.

Meta: Led the design and prototyping of a voice experience for their Portal device. Outcomes included a deep assessment of the concept's desirability, usability recommendations and a strategic roadmap for further development.

Jason Wolfe

Lead Platform Engineer

Jason Wolfe serves as the Lead Platform Engineer for Thoughtworks, bringing more than a decade of experience in enterprise modernization, infrastructure architecture, and platform engineering across highly regulated industries. As the Commonwealth establishes the Data Commons Collaborative (DCC) as the secure, scalable foundation of the Massachusetts AI Hub, Jason will oversee the end-to-end engineering of the hybrid cloud environment—ensuring that it is automated, resilient, and built for

extensibility across agencies and research partners. His work will focus on implementing the RFP's requirements for infrastructure-as-code, security, observability, and zero-trust operations, enabling the DCC to scale responsibly and sustainably.

At Thoughtworks, Jason serves as a Lead Consultant and Infrastructure Architect, specializing in platform enablement and modernization for data-intensive organizations. He has led several large-scale transformations, including data mesh for pharmaceutical clients, where he designed and automated cloud-native platforms that support both operational workloads and AI-driven research. His expertise spans AWS, Terraform, Kubernetes, and CI/CD automation, allowing him to implement multi-environment, auditable deployments that align with Section 6.0 of the DCC RFP (Security, Governance & Compliance). Jason's leadership philosophy blends engineering rigor with collaboration—enabling cross-functional teams to deploy infrastructure safely, rapidly, and at scale.

Before joining Thoughtworks, Jason served in senior technology roles across the private sector, including as Global Engineering Director at Sungard Availability Services and as a transformation consultant for Norwegian Cruise Line Holdings. He has advised C-level clients on modernization roadmaps and implemented cloud landing zones that balance performance, compliance, and cost optimization. With a PhD in Management Science and an MBA in International Business, Jason brings both the technical acumen and organizational insight required to operationalize the Commonwealth's long-term AI infrastructure vision. His proven ability to align technical architecture with mission outcomes ensures the DCC platform will deliver high availability, transparency, and trust for every participating institution.

6. Relevant experience & References

Client	United States Department of Veterans Affairs (VA)

Case Study Summary

Built an open-source–based digital notifications platform (VANotify) and advanced a data-as-product program (MPR data product on Azure with a Collibra catalog) to modernize citizen communications and enable trustworthy analytics at scale.

Thoughtworks Approach

Engineered VANotify on AWS GovCloud with Infrastructure as Code and secure CI/CD; introduced Data Mesh concepts; implemented a Collibra-backed catalog to improve discoverability, lineage, and policy enforcement.

Benefits & Outcomes

1.5M+ digital notifications with ~70% open rate and 200k+ vaccination-planning messages; MPR processing time reduced by 88%, with a governed catalog reducing time-to-insight and improving reuse across domains.

Reference Details: Marian Adly marian.adly@va.gov (630) 677-2718

Client	Gilead Sciences

Case Study Summary

Partnered on a multi-year transformation to accelerate "molecule to market" timelines by modernizing Gilead's data ecosystem. We implemented a Data Mesh architecture to democratize data access and built a self-service cloud platform to automate infrastructure provisioning, enabling secure, governed, and scalable AI/ML innovation across R&D, manufacturing, and commercial domains.

Thoughtworks Approach

We employed a two-pronged strategy focused on data architecture and platform engineering: 1. Data Mesh Strategy & Implementation: Guided Gilead’s adoption of a decentralized Data Mesh architecture on AWS. This involved: * Establishing federated computational governance to empower domain-owned data products (R&D, Supply Chain, Commercial) while enforcing central policies using tools like AWS Lake Formation

and Collibra. * Treating Data as a Product, ensuring all assets are discoverable, trustworthy, and secure. * Shifting the organization to cross-functional, product-oriented teams to accelerate value delivery. 2. Cloud Platform Automation: Engineered a self-service Cloud Experience Portal to eliminate a critical operational bottleneck. This solution: * Automated the entire lifecycle of AWS EC2 instances (provisioning, modification, termination). * Embedded compliance-as-code, enforcing security baselines, hardened AMIs, and KMS encryption by default. * Utilized a modern stack (Terraform, Kubernetes, GitHub Actions, AWS Lambda) for a resilient, auditable, and automated platform.

Benefits & Outcomes

Delivered significant, measurable improvements in speed, efficiency, and compliance: * Drastically reduced time to market for cloud infrastructure, cutting the provisioning time for compliant EC2 instances from 5 business days to just 30 minutes. * Established an enterprise-wide federated data governance model, improving data trust and accelerating responsible, cross-domain analytics and AI/ML enablement. * Increased operational efficiency and enhanced security by default, freeing up engineering teams from manual ticketing and compliance checks. * Created a scalable, self-service foundation that empowers teams to innovate faster while adhering to rigorous security and regulatory standards. See: https://www.thoughtworks.com/about-us/events/webinars/data-mesh-in-action https://www.thoughtworks.com/insights/blog/data-engineering/data-mesh-real-examples-and-lessons-learned

Reference Details: Upon request, 3 others provided.

Client	State of Illinois - BEACON

Case Study Summary. Launched a single, public entry point for families and caregivers to discover and connect with state-funded and community resources for youth behavioral health, integrating information and workflows across six Illinois agencies (DHS, DCFS, ISBE, IDPH, HFS, IDJJ). Delivered a rapid, MVP-first program that stood up core capabilities in approximately six months, then iterated to harden, expand for statewide launch. BEACON’s structured program metadata and discovery patterns align with DCC’s data catalog and FAIR principles for findability and reuse.

Thoughtworks Approach: Cloud-native architecture on Google Cloud (GCP) with infrastructure-as-code, identity and access controls, logging, and auditability to meet public-sector security and compliance needs. API-first, modular services to support guided intake, consent-aware information sharing, and role-segmented workflows for parents/guardians, resource coordinators, clinical specialists, and agency staff. Human-centered delivery using agile, thin-slice increments focused on a usable MVP, continuous stakeholder demos, and progressive hardening for statewide use. Operational transparency and analytics leveraging Looker/BigQuery for program dashboards and case progress monitoring to inform policy and operations.

Benefits & Outcomes: One “no-wrong-door” portal at statewide scale, simplifying a previously fragmented, multi-agency journey for Illinois families seeking behavioral health services for youth. Six-month MVP to production launch cadence, demonstrating speed-to-value while preserving security, governance, and accessibility requirements for the public sector. Security and governance by design, including RBAC, MFA/IdP integration, encryption in transit and at rest, and long-retention audit logs aligned to government standards on GCP.

Reference Details: Dr. Dana Weiner dana.a.weiner@illinois.gov (312) 339 8884

Client	State of Oklahoma

Case Study Summary: Thoughtworks partnered with the State of Oklahoma across OMES and OESC to modernize core systems for integrity, access, and fraud prevention—standing up GCP-based platform enablement, automating privacy-preserving obfuscation for non-production environments, delivering operational dashboards, and integrating cross-agency identity verification via VerifyOK to streamline benefits delivery and reduce fraud.

Thoughtworks Approach: Implemented a secure, cloud-native foundation on GCP, aligned to Google’s BeyondCorp principles to strengthen identity- and context-aware

access across agencies. Automated privacy-preserving data obfuscation for pre-production using Google DLP, enabling safer testing and faster delivery while protecting sensitive information. Advanced API-led modernization and delivered operational dashboards to eliminate manual reporting steps and provide near-real-time visibility into program performance. Integrated cross-agency identity verification by partnering with Idemia's SaaS for VerifyOK, orchestrated through OMES to drive reuse and uniform controls statewide. Responsible and Ethical Operation Readiness: While Oklahoma's engagement focused on security, identity and data protection, the same patterns enable the RFP's fairness, auditing, and synthetic-data capabilities (e.g., logging for traceability; controlled access; data provenance), which our response operationalizes with tools like **AIF360**, **Fairlearn**, and **SHAP.**
Benefits & Outcomes: Eliminated manual reporting steps and improved operational visibility with dashboards, accelerating decision-making and reducing effort across teams. Enabled secure, production-like testing by automating non-prod data obfuscation with Google DLP, improving delivery velocity while protecting sensitive data. Deployed VerifyOK as a flagship, cross-agency identity verification product—leveraged by up to 178 agencies via OMES—to harden integrity, prevent fraud, and increase reuse of a statewide capability
Reference Details: Joshua Walcher joshua.walcher@okdhs.org (405) 219-5363

Other Selected Client Stories

Client	Case Study Summary	Thoughtworks Approach	Benefits & Outcomes
State of Oklahoma (OMES, OESC)	Modernized core systems to improve integrity and access: GCP-based platform enablement, privacy-preserving obfuscation for pre-prod, operational dashboards, and	Delivered BeyondCorp-aligned security, automated obfuscation with Google DLP, API-led modernization and dashboards; partnered with Idemia SaaS for VerifyOK and	Eliminated manual reporting steps and enabled secure pre-prod testing; VerifyOK deployed as a flagship product leveraged by up to 178 agencies, improving fraud

Client	Case Study Summary	Thoughtworks Approach	Benefits & Outcomes
	VerifyOK identity verification to reduce fraud and streamline benefits across agencies.	integrated identity verification across agencies via OMES.	prevention and cross-agency reuse.
Sport Australia	Co-created SportAUS Connect, an integration and interoperability platform providing digital identity and shared foundations; delivered a 12-week analytics prototype to inform evidence-based infrastructure investments across jurisdictions.	Introduced agile/DevOps ways of working; delivered on Microsoft Azure leveraging RedHat Keycloak SSO, CI/CD, and reusable platform components to onboard future products faster.	A reusable, evolvable platform and digital identity service enabling partners to onboard more products, plus an analytics prototype demonstrating actionable spatial insights for national investment decisions.
ITV	Co-developed a cloud Data Mesh on AWS and Databricks to operationalize value-first data strategy, from thin slices through to activation, accelerating audience targeting and commercial use cases across domains.	Built a self-serve mesh platform, domain data products, and a CDP bridge for activation; embedded federated governance and standards to ensure interoperability and auditability.	Reduced audience campaign setup from three months to three minutes; improved commercial effectiveness by ~95%; reduced data product provisioning from ~3 weeks to hours; 20+ data products in production.
Saxo Bank	Enabled data democratization with	Evaluated and integrated tools (e.g.,	MVP for lineage and glossary in five

Client	Case Study Summary	Thoughtworks Approach	Benefits & Outcomes
	"Data Workbench," a self-service data catalog, business glossary, lineage, and data-quality capabilities aligned to Data Mesh and open banking goals.	LinkedIn DataHub + DQ framework) into Saxo's streaming-first ecosystem; defined personas, standards, and onboarding flows to scale governed reuse.	months; improved trust and discoverability; open-source contributions to DataHub increased visibility and accelerated governance adoption.
Gilead Sciences	Planned and initiated an enterprise Data Mesh and AI platform transformation on AWS to scale domain-owned data products and responsibly enable AI/ML across R&D, manufacturing, supply chain, and commercial.	Co-developed data/AI strategy, operating model, and platform roadmaps; applied product thinking with thin-slice delivery patterns to accelerate adoption and AI-readiness.	Enhanced cross-domain data access and governance guardrails, accelerating responsible AI enablement and time-to-insight across regulated functions.
Bayer	Built PRINCE, an AWS-based preclinical data platform unifying structured and unstructured data (including ~17,000 study reports); later added a GenAI, RAG-powered chatbot that won Bayer's 2025 GenAI "Best Technical Implementation."	Engineered a secure, compliant data platform with compliance-as-code; delivered a domain-specific RAG chatbot on Bayer's myGenAssist, with observability and source-cited responses for trust and auditability.	Faster, more reliable research and decision support; reduced rework and accelerated preclinical discovery; recognized internally with a GenAI award for technical excellence.

7. Thoughtworks partnerships and accelerators

Thoughtworks believes a successful, sustainable, and scalable Data Commons Collaborative (DCC) must be built on proven, best-in-class technology that supports open standards while meeting rigorous governance requirements. Our delivery strategy is fundamentally underpinned by strategic partnerships with Databricks and Google Cloud (GCP), ensuring we bring a compelling, comprehensive, and low-risk solution to the Massachusetts AI Hub. Our proposed pilot leverages the Databricks platform as a strategic accelerator, enabling us to deliver value to the Commonwealth's innovation community with speed and efficiency. This approach is aligned with the long-term vision of creating a sovereign, open-source foundation as Databricks itself is built upon powerful open technologies, including Apache Spark for data processing and MLflow for MLOps. By validating the core data and machine learning workflows within this managed environment, we establish a clear and low-risk pathway to a fully open architecture. The skills, patterns, and data assets developed during the pilot can be seamlessly transitioned to a self-managed stack post-pilot, where proprietary components like the Unity Catalog could be replaced with open alternatives such as DataHub, ensuring the Commonwealth retains ultimate control over its technology destiny without sacrificing initial speed-to-market

Partner	Technology	Capability Mapping
Databricks	Databricks Lakehouse & Unity Catalog, Managed MLFlow	Data Curation, Metadata Management, RBAC, Data Lineage, Scalable Storage, AI Fairness Integration
Cloud Platform (AWS, Azure, or GCP)	CSP Native Services	Scalable Compute (GPU/CPU), Model Serving (API), Security/Compliance, Overall Cloud Foundation.

Databricks as a unified data foundation

Thoughtworks is proud to partner with Databricks, a leader in data analytics and AI and leverage their cutting-edge solutions to align perfectly with Thoughtworks' groundbreaking innovation in data platform modernization.

Our partnership leverages the Databricks Data Intelligence Platform and the Unity Catalog as the core data layer for the DCC. This architecture directly addresses the need for centralized, governed data access.

- **Centralized Governance (Unity Catalog):** Unity Catalog is a key component in order to achieve granular Role-Based Access Control (RBAC), Auditing, and Dataset Discovery requirements. It will serve as the single source of truth for all data, models, and metadata, ensuring that Curators, Admins, and Researchers receive appropriate

access, audit trails, and data lineage. This ensures centralized governance from a single platform.

- **Data Ingestion & Curation:** We will utilize Databricks' native Lakeflow Connect and Auto Loader tools to ingest Commonwealth and user-submitted data. This greatly limits the need to build custom ingestion pipelines from scratch, and hence reducing development cost and accelerating the path to a Bronze (Raw) and Silver (Curated) data layer.
- **Scalability and Openness:** Databricks provides the necessary scale for large datasets and adheres to the open Delta Lake format, ensuring the DCC is extensible and avoids vendor lock-in, aligning with the Hub's long-term vision.

Databricks' fully managed MLflow MLOps Platform is central to connecting the Databricks Lakehouse with the Vertex AI serving layer, providing the necessary governance for the entire model lifecycle.

- **Model Lineage and Versioning:** MLflow's tracking and registry features provide the audit trail for models, ensuring every deployed model is traceable back to the exact code, parameters, and dataset version used to train it.
- **AI Fairness Integration:** MLflow natively integrates with key ethical frameworks and tools (e.g., Fairlearn, AIF360) that we will deploy. This enables the team to embed bias detection and auditing directly into the model development workflow ensuring a transparent and accountable AI Fairness and Bias Framework.
- **Unified View:** MLflow acts as the central control plane for all experimental results, allowing curators and administrators to compare model performance, review fairness metrics, and approve models for deployment in a single, auditable interface.

Cloud platform as core infrastructure and AI platform (AWS, Azure, or GCP)

Thoughtworks is a Premier Partner with all three major cloud service providers (CSPs) - AWS, Azure, and GCP and our partnership brings industry-leading platform and Thoughtworks' world-class delivery capabilities to the Commonwealth of Massachusetts ensuring innovation and scale for DCC.

If Thoughtworks is selected as the chosen partner for the DCC initiative, during the discovery phase of the engagement, we will partner with The Commonwealth and conduct a thorough analysis to determine the CSP that best suits the needs of DCC. Then, we will leverage the respective services of the CSP as the foundation for our core infrastructure and AI components like compute, storage, application deployment layer, and model deployment layer.

Regardless of the final cloud environment chosen by MassTech, our solution is designed to leverage native cloud services for optimized performance and compliance:

- **Model Serving:** We will utilize the native cloud MLOps service (e.g., Google's Vertex AI, AWS's SageMaker, or Azure's Azure ML) for the final production deployment. This ensures the delivery of scalable, low-latency **API endpoints** required for the public website. Another option to consider here is to leverage Databricks for full MLOps lifecycle management as well.
- **Security and Compliance:** The native cloud platform's security services provide the robust infrastructure for **logging, encryption (at rest and in transit), and audit trails** required by the RFP's stringent Security and Governance section.
- **High-Performance Compute:** We will provision native cloud-based **GPU/CPU environments** for all synthetic data generation and model training tasks, optimizing cost and performance.

8. Assumptions

In order to achieve pilot release in published not-to-exceed budget the Thoughtworks team states the following assumptions:

- Pilot audience identified during initial 2 weeks of engagement
- DCC stakeholders available for working sessions / interviews as required by program plan
- Pilot audience users available for feedback sessions
- Pilot audience and stakeholders prioritization of must have/nice to have features and data sets can be accomplished inside the published budget/timeline
- Tool licensing/any DCC contracting to happen within timelines in the program plan
- Prioritized data sets made available via documented integration methods
- Thoughtworks team members onboarding to take no more than 3 business days from start date
- Thoughtworks team members will have access to all necessary documentation and environments
- While scope might include restricted access to certain datasets - the team assumes restrictions will be managed by the identity/access management solution
- The team assumes datasets will be certified de-identified and therefore not subject to regulatory restrictions above and the identity/access management solution
- Acceptance: Criteria will trace to the enumerated features and NFRs. External certifications (e.g., Section 508 audits, red-team/pen tests) and continuous monitoring/SOC services are excluded unless added by change order.
- The solution will interoperate with MGHPCC/AICR and the Commonwealth's environment. Thoughtworks will design and configure; the Client will procure the selected infrastructure/services. Procurement or quota delays are outside Thoughtworks' control and may impact schedule.
- FAIR/linked-data standards (e.g., W3C/DCAT/Schema.org) will be applied pragmatically. If later-mandated standards require rework, they will be handled via change control.
- In absence of mandated frameworks, we align to Commonwealth norms and industry best practices (e.g., NIST). Additional compliance artifacts beyond standard delivery (e.g., SSPs, POA&Ms, penetration tests) are excluded unless explicitly requested.

- Third-party services: Licenses/quotas (e.g., Snowflake compute, MGHPCC/AICR capacity, scanning tools, commercial APIs) are assumed to be client-procured and funded unless explicitly included in our pricing/cost table.
- The solution will meet WCAG 2.1 AA, third-party accessibility certification/audits are out of scope unless specifically requested and funded.
- Data storage not expected to exceed 1 TB through the two phases, with a maximum processing of 100 GB during the pilot.
- Pilot costs calculated based on the assumption of 25 users. Additional users would incur additional cost.

Mike O'Brien

Client Principal

mike.obrien@thoughtworks.com

www.thoughtworks.com

© 2025 Thoughtworks Commercial in Confidence

LEGEND:

Formula cell

Data entry cell

This section is for post-contract invoicing only. Do not complete with proposal.

Line Number	Budgeting															
2	Applicant Information															
3	Applicant: Thoughtworks			Procurement No.:2026-JAII-01												
4	Budget Period:			Title of Proposed Project: AI Hub Data Commo												
5	Address: 200 E Randolph St 25th Floor, Chicago, IL 60601															
6																
7																
8				MTC Funding Requested: $6,687,072												
9				Total Project Cost:$6,827,072												
10	Cost Elements			Total Project	Invoice 1		Invoice 2		Invoice 3		Invoice 4		Cumulative			
11					Period: MM/DD/YYYY-MM/DD/YYYY		Period: MM/DD/YYYY-MM/DD/YYYY		Period: MM/DD/YYYY-MM/DD/YYYY		Period: MM/DD/YYYY-MM/DD/YYYY		Period: MM/DD/YYYY-MM/DD/YYYY			
12						Amount		Amount		Amount		Amount	Expended		Remaining	
13	I. Direct Labor															
14	*name/title*	*HRS*	*Rate*		*HRS*		*HRS*		*HRS*		*HRS*					
15	Data Architect	672	$344	$231,168		$ -		$ -		$ -		$ -	$ -		$	231,168.00
	ML Engineer	2,016.00	$341	$687,900		$ -		$ -		$ -		$ -	$ -		$	687,899.52
16	Data Engineer	2,864.00	$310	$888,413		$ -		$ -		$ -		$ -	$ -		$	888,412.80
	Full-Stack Engineer	2,264.00	$282	$638,448		$ -		$ -		$ -		$ -	$ -		$	638,448.00
	Full-Stack Engineer	2,864.00	$243	$695,952		$ -		$ -		$ -		$ -	$ -		$	695,952.00
	UI/UX	2,264.00	$273	$618,072		$ -		$ -		$ -		$ -	$ -		$	618,072.00
	Product Manager/Soln Owner	2,016.00	$293	$590,688		$ -		$ -		$ -		$ -	$ -		$	590,688.00
17	QA	2,264.00	$256	$579,584		$ -		$ -		$ -		$ -	$ -		$	579,584.00
18	PM/DP	2,264.00	$282	$638,448		$ -		$ -		$ -		$ -	$ -		$	638,448.00
19	Infra/Security	2,016.00	$282	$568,512		$ -		$ -		$ -		$ -	$ -		$	568,512.00
20	UI Developer	2,016.00	$243	$489,888		$ -		$ -		$ -		$ -	$ -		$	489,888.00
21	Total Direct Labor			$6,627,072	0	$ -	0	$ -	0	$ -	0	$ -	$ -		$	6,627,072.32
22																
23	II. Direct Labor Fringe Cost	Rate (%):		$0		$ -		$ -		$ -		$ -	$ -		$	-
24																
25	II. Subcontractors/Consultants/Contracted Staffing															
26	*Firm name/title or role/type of services to be provided*	*HRS*	*Rate*		*HRS*		*HRS*		*HRS*		*HRS*					
27				$0		$ -		$ -		$ -		$ -	$ -		$	-
28				$0		$ -		$ -		$ -		$ -	$ -		$	-
29				$0		$ -		$ -		$ -		$ -	$ -		$	-
30				$0		$ -		$ -		$ -		$ -	$ -		$	-
31				$0	5	$ -		$ -		$ -		$ -	$ -		$	-
32				$0		$ -		$ -		$ -		$ -	$ -		$	-
33	Total Subcontractors/Consultants			$0	5	$ -	0	$ -	0	$ -	0	$ -	$ -		$	-
34																
35	III. Direct Materials (provide description and costs basis)															
36				$0									$0		$	0
37				$0									$0		$	0
38				$0									$0		$	0
39				$0									$0		$	0
40				$0									$0		$	0
41				$0									$0		$	0
42				$0									$0		$	0
43	Total Direct Materials			$0		$0		$0		$0		$0	$0		$	0
44																
45	IV. Travel												$0		$	0
46																
47	V. Other Direct Costs *(list type and description such as venue rental costs and cost basis)*															
48				$0									$ -		$	-
49				$0									$ -		$	-
50				$0									$ -		$	-
51				$0									$ -		$	-
52				$0									$ -		$	-
53	Total Other Direct Costs			$0		$ -		$ -		$ -		$ -	$ -		$	-
54																
55	VI. General & Administrative Expense/Overhead	Rate (%):		$0		$ -		$0		$0		$0	$ -		$	0
56																
57	**Total PROJECT COSTS**			**$6,627,072**		**$0**		**$0**		**$0**		**$0**	**$0**		**$**	**6,627,072**